时态主题模型方法及应用研究综述

桂小庆 张 俊 张晓民 于鹏飞
(大连海事大学信息科学技术学院 大连 116026)

摘 要 互联网技术的飞速发展使得数据的规模达到了空前的水平,人们从海量数据中获取有价值的信息变得越来越困难。主题模型是近年来计算机领域出现的一种新的概率模型,在自然语言处理、文本挖掘以及信息检索等领域都有很广泛的应用。基于主题模型的主题追踪技术和时态分析技术可以帮助人们从海量数据中快速找到感兴趣的内容,时态主题模型逐渐成为计算机科学领域的一个研究热点。首先,介绍主题模型以及时态主题模型的基本概念;然后,对各种时态主题模型进行分类,介绍了几种具有代表性的时态主题模型,分析比较了各种主题模型的优缺点;接着,分析了时态主题模型在社交媒体、学术文献和数据社区中的应用;最后,对时态主题模型未来的发展趋势进行了探讨。

关键词 时态,主题模型,时态主题模型
中图法分类号 TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.02.005

Survey on Temporal Topic Model Methods and Application

GUI Xiao-qing ZHANG Jun ZHANG Xiao-min YU Peng-fei
(College of Information Science and Technology, Dalian Maritime University, Dalian 116026, China)

Abstract With the fast development of Internet, the data has reached an unprecedented scale. However, it is becoming more and more difficult to get valuable information from mass data. Topic model is a new probabilistic model which has been widely applied in natural language processing, text mining, information retrieval and other fields in recent years. The technology of topic detecting and temporal analysis can help users focus on interested information. Temporal topic model has gradually become a hot research topic in the field of computer science. Therefore, temporal topic model and its application were investigated in detail in this paper. Firstly, the basic knowledge of topic model and temporal topic model were introduced. Secondly, temporal models were categorized into several types, representative models were discussed and their advantages and disadvantages were also analyzed. Thirdly, the applications of temporal models were summarized in several fields. Finally, the future development trends of temporal topic models were presented.

Keywords Temporality, Topic model, Temporal topic model

1 引言

随着信息技术的发展,人类正以史无前例的速度生产新的数据,全球互联网中的数据量以指数级的方式增长。为了便于理解大规模的数据,总希望能用一个简短的描述或主题来代表一个数据集的特征信息。例如对于文本数据,就是抽取出若干个主题来描述文本数据集。因此,研究人员于2003年第一次显式地提出了一种可以发现文档集中隐含主题的统计模型[1],即主题模型(Topic Model)。

主题模型是一种对文本隐含主题进行建模的方法,通过将高维度的词的集合映射到低维度的主题空间上来对目标数据进行降维,建立简洁的表示。同时,通过对该模型降维获得的低维空间的表示具有很好的可解读性,可以较容易地转化为人们可理解的主题信息,为挖掘文本中的隐含语义提供了可能,并且能够自动寻找海量数据中的语义主题,有助于人们理解文本集中所讨论的内容。

在实际应用中,简单的主题模型已难以满足日益增长的对特定数据的快速获取的需求。随着在线媒体的广泛应用,文本具有明显的时间属性,现实应用中经常需要结合时间信息来解释主题[2]。若文本具有时间属性,那么就很有必要识别在不同的时间下的不同的主题,并分析主题基于时间的演化,因此传统的方法在对现有主题进行检测与追踪和挖掘该类文本时的可用性不足。时态主题模型可以充分利用文本中的时态信息,分析文本中的主题随时间演化的规律,从而帮助用户追踪感兴趣的话题。由于主题模型具有良好的扩展性和广泛的应用前景,近几年来吸引了越来越多研究人员的关注,

到稿日期:2016-01-02 返修日期:2016-05-02 本文受国家自然科学基金(61073057,61370070)资助。
桂小庆(1990—),女,硕士生,主要研究方向为数据库与信息检索,E-mail:guixiaoqingdmu@163.com;张 俊(1971—),男,博士,教授,主要研究方向为数据库与信息检索;张晓民(1991—),男,硕士生,主要研究方向为数据库与信息检索;于鹏飞(1991—),男,硕士生,主要研究方向为数据库与信息检索。

并已被广泛地应用于信息处理的各种任务中。近年来，基于时态的主题模型扩展成为主题模型的热点研究方向，获得了较多的研究成果。

本文第 2 节从 LDA 主题模型、时态主题模型、主题模型的应用等几个方面介绍主题模型的发展；第 3 节介绍时态主题模型的基本知识，并从简单时态主题模型和扩展的时态主题模型两个方面对时态主题模型的研究现状作简单分析；第 4 节从社交媒体、学术文献、数据社区等 3 个方面介绍时态主题模型的应用；最后对时态主题模型的研究和发展趋势进行了展望。

2　主题模型发展概述

主题模型是一种语言模型，是在机器学习和自然语言处理等领域中用于在一系列文档中发现隐含主题的一种统计模型，同时允许从简单的无秩序的特征数据中抽取出主要的模式，以实现文本语义挖掘[1]。主题模型还以通过使用贝叶斯统计和机器学习方法发现非标签文档的潜在语义内容，并利用这些潜在语义对文档集的未来特性进行预测。主题模型是文档的一种生成模型：根据主题模型所指定的一组概率程序创建出一个新的文档。首先选择一个主题的概率分布，然后根据这个概率分布，每次随机地从中选出一个主题，再根据这个主题在单词上的概率分布，生成这个文档的一个个单词，这样就可以产生一个新的文档。同时根据该概率生成模型对已有的文档集进行反向操作，统计推断这个文档集具体的主题概率分布以及每个主题在字词上的概率分布，最终得到主题信息。其过程如图 1 所示。

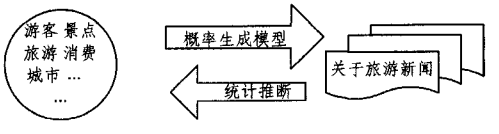

图 1　主题模型示意图

主题模型的思想最早来源于潜在语义分析（Latent Semantic Analysis，LSA）[3]，它由 Papadimitriou 于 1998 年首次提出，随后在 1999 年 Tomas Hofmann 提出了概率潜在语义分析（Probabilistic Latent Semantic Analysis，PLSA）[4]，这也是最早提出的概率生成主题模型。在 LSI 和 PLSI 的基础上，2003 年 David Blei 等人提出潜在狄利克雷分布（Latent Dirichlet Allocation，LDA）模型[1]。

2.1　LDA 主题模型概述

LDA 是由 David Blei 等人于 2003 年提出的一种概率主题模型[1]，作为当前最常用的一种主题产生工具，其由于具有良好的数学基础和灵活的扩展性，正受到越来越多研究者的高度关注，并被广泛地应用于各种研究领域。

LDA 模型将文本映射到主题空间，即认为一个文本由若干主题随机组成，从而获得文本间的关系。LDA 模型是一种生成模型，建立在简单的概率抽样基础上，描述了在（随机）隐含变量之上生成文本的过程，即 LDA 是一个“文本—主题—词”的三层贝叶斯产生式模型。LDA 的目的是发现隐藏在大量文档之后的主题，通过观察已知的文档去获取或者推出隐藏在它们之后的主题结构[5]。

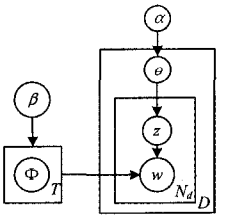

图 2　LDA 主题模型

图 2 是一个简单的 LDA 图模型，该模型有 3 层表述，参数 α 和 β 是整个文档库层面的表述，用以在概率生成过程中生成整个文档集合。变量 θ_d 是单个文档层面的表述，用于生成某个文档。变量 z_{dn} 和 w_{dn} 是词层面的表述，用以生成每个文档中的每个词语。

LDA 模型产生文本的过程如下：

（1）对于每个主题 Z，从具有参数 β 的 Dirichlet 分布中选取一个多项式分布 Φ_z，Φ_z 是指在主题 Z 中模型化特征的相对频率；

（2）对于每篇文章 d，从参数为 α 的 Dirichlet 分布中选取一个多项式分布 θ_d，θ_d 是指在文档 d 中模型化主题的相对频率；

（3）对于文章中的每个词 W，从多项式分布 θ_d 中选取一个主题 $z\in\{1,\cdots,K\}$，结合多项式 Φ_z 来选择词 W。

由此可知，生成一个文档的概率过程为：

$$p(D|\alpha,\beta)=\sum_{d=1}^{M}\int P(\theta_d|\alpha)\Big[\sum_{n=1}^{N_d}\sum_{z_{dn}}P(z_{dn}|\theta_d)P(w_{dn}|z_{dn},\beta)\Big]d^k\theta_d$$

2.2　时态主题模型概述

目前大部分的主题模型都是一种静态主题模型，即特点是：1）数据的统计量一旦获取就不会轻易改变，适用于静态的数据；2）一旦得到模型，就不会再有信息的变更；3）不会使用学习算法来使得模型发生变化。静态主题模型考察没有时间维的文本集的主题分布情况，通过对文本的建模，可以对文本进行主题分类，判断文档相似度等[6]。静态主题模型假设文档是“Bags of Words”，在主题发现方面具有明显优势，可以实现文档集合上隐含语义的挖掘。静态主题模型针对静态的文档集，模型的建立与时间无关，即假设主题不随时间变化。然而现实生活中，文档集都是动态变化的。在文本集中，每个文本除了文本内容特征外，一个非常关键的要素就是时间。大多文本集都随时间不断变化，并且不同时间段内的主题也随之改变，因此当需要对动态变化的数据进行分析和主题追踪时，静态主题模型是难以满足这种需求的。因此，2006 年 Blei 等人首次提出了动态主题模型（Dynamic Topic Model）[7]来解决此类问题。研究人员在主题建模时通过引入时间维度，从而跟踪分析主题随时间动态变化的状况（新产生、变强或变弱），或者跟踪主题的内容变化情况，即主题内的词分布的变化情况（新增加、变强或变弱）[8]。时态主题模型相对于静态主题模型而言，能更及时地响应用户的相关信息的变化。随着用户兴趣的变化，模型通过学习或者与系统的交互，可以获知用户兴趣的变化并增量式地更新主题模型。

2.3 主题模型应用概述

随着主题模型的推出，主题模型方法几乎被应用到了所有的文本挖掘和信息处理领域中。由于 LDA 主题模型具有良好的扩展性，主题模型的相关研究工作大多是对 LDA 模型进行扩展，或是将 LDA 模型作为某个概率模型的一个部件，被直接或扩展使用在自然语言处理的众多任务中[9]。静态主题模型中，例如 AT(Author-Topic Model)[10]利用主题对文章作者和文章内容建模，主题分布受到作者分布的影响。CTM(Correlated Topic Model)[11]不仅考虑了文档之间的关系，还引入了主题间的关系。对于时态主题模型，以 ATT(Author-Topic-Time Model)[12]为例，通过引入文本集中所包含的时间戳信息，不仅能够获取用户信息相对时间的变化，同时还实现了对主题演化过程中主题趋势的分析。

目前主题模型已经被成功地应用到情感分析、学术文献挖掘、社交媒体分析、网络结构化数据挖掘、时态文本流分析等诸多应用中。因此，以 LDA 为基础的主题模型在国内外获得了普遍认可并且成为了广大研究人员密切关注的研究领域。表 1 描述了目前一些主流的主题模型的提出时间、提出作者及其应用领域。

表 1 主题模型及其应用汇总

时间	作者	模型	适用领域
2003	Blei 等[1]	LDA	情感分析、学术文献挖掘、社交媒体主题挖掘
2008	Blei 等[13]	sLDA (Supervised LDA)	情感分析、文本分类
2006	Blei 等[11]	CTM(Correlated Topic Model)	新闻及社交媒体中对主题相关性的建模
2004	Steyvers 等[10]	AT (Author-Topic Model)	文献挖掘中对文献作者的建模
2005	McCallum 等[14]	ART (Author-Recipient-Topic)	文献挖掘、主题聚类
2008	Nallapati 等[15]	Link-PLSA-LDA	文献挖掘、网络结构数据的挖掘
2011	Nasir Naveed 等[12]	ATT (Author-Topic-Time model)	文献挖掘、时态文本流
2009	Ramage 等[16]	Labeled LDA	文本分类
2006	Wang 等[17]	TOT (Topic Over Time)	时态文本流
2006	Blei 等[7]	DTM (Dynamic Topic-Model)	时态文本流
2008	AlSumait 等[18]	OLDA	社交媒体主题挖掘、推荐系统
2011	Wayne Xin Zhao 等[19]	Tweet-LDA	社交媒体主题挖掘
2009	Blei 等[20]	RTM (Relational Topic Model)	社区发现
2013	Daifeng Li[21]	DCTM (Dynamic Community TopicModel)	社区发现

3 时态主题模型

近年来时态主题模型已得到了越来越多的关注、研究和应用。时态主题模型的种类较多，并且实现方法、功能特点及适用领域各异，本节将对时态主题模型进行较为详细的分类总结与分析。

3.1 简单的时态主题模型

简单的时态主题模型将时间信息结合到主题模型中，得到的主题表现出随时间演化的趋势，而且这类模型大多以 LDA 的扩展为主。这类模型从数据时间的划分角度可以分为两大类[22]：1)将时间视为随机变量，从而进行连续时间的建模；2)把时间离散化为一系列时间戳，进而对离散化的时间点构建动态贝叶斯网络。

3.1.1 Topic Over Time(TOT)模型

2006 年 Wang & McCallum[17]提出了 TOT 模型，它是第一类模型的典型代表，也是 LDA 型的一种主题抽取模型，通过在 LDA 模型中引入时间因素构建而成，实现简单方便。

TOT 模型认为主题的发现不仅仅受到词共同出现的频率的影响，还受到时间变化的影响，例如近几年某个主题比较热门，在整个数据集中占的比重自然就会大一些，若未来几年研究者对这个主题的研究比较深入，或是难以再深入研究，那么关注这个主题的人自然会越来越少，从而这个主题在数据集中的分布比重就要下降。

TOT 模型不仅能够获取低维的数据结构，也能分析这些数据随着时间的结构变化。在此之前的一些研究都依赖于 Markov 推理或忽略了数据的时间关系，而与此不同的是，TOT 模型采用了非马尔科夫方法对主题演化趋势进行建模，是一个基于 LDA 的生成过程，它将时间连带主题发现中词汇的共同出现类型进行模型化。对于每一篇生成文档，词的共现以及文档的时间戳共同影响主题的混合分布。因此，一个主题中词的共现及基于时间的关联改变对研究具有重要的意义。TOT 模型是对具有时间属性的文档中的时间戳和词的一种生成模型，TOT 模型的生成过程如下[17]：

(1)对于每个主题 Z，从具有参数 β 的 Dirichlet 分布中选取一个多项式分布 φ_z；

(2)对于每篇文档 d，从参数为 α 的 Dirichlet 分布中选取一个多项式分 θ_d 布；对于文档中的每一个单词 w_{di}：

1)从多项式分布 θ_d 中生成一个主题 z_{di}；

2)从多项式分布 $\phi_{z_{di}}$ 中生成一个单词 w_{di}；

3)从 beta 分布 $\psi_{z_{di}}$ 中生成一个时间戳 t_{di}。

由上述可知，其生成过程与 LDA 类似，只是每个词多了一个时间属性。TOT 的贝叶斯图模型如图 3 所示，由图可以看出，对于每一篇文档，根据词的 β 分布抽样出时间戳 t，并且每篇文档中的所有的词具有相同的时间戳。

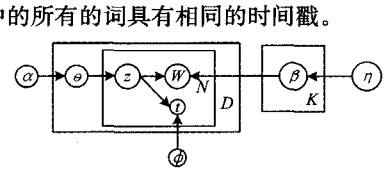

图 3 TOT 图模型

尽管 TOT 模型考虑了文本的时间信息，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，忽略了主题内容的变化[23]。

(2)对于随时间变化较为频繁以及规模较大的文本集而言，使用 TOT 模型将消耗大量的计算资源和内存资源。

3.1.2 Dynamic Topic Model(DTM)模型

由于 TOT 模型无法解决对动态变化数据的分析和主题

内容追踪，2006 年 Blei 等人提出了动态主题模型（Dynamic Topic Model，DTM）[7] 用于解决此类问题。DTM 是第二类模型的代表，被证明是一个能够准确描述潜在主题及其动态变化的强有力的工具。

动态主题模型利用时间序列分析方法生成每一片段的主题结构。其主要思想是：首先按照时间顺序将数据划分为若干个片段，每一个片段按照静态主题模型的思想建模，最终形成主题随时间的演化。根据这种情况，在动态主题模型中，数据集按照时间戳的设置被分为多个子集合，每个单词集合被赋予一个特定的时间戳。同时，假设主题数量是固定的，即在每个时间戳中，文本都是由固定的主题数量的 LDA 模型生成。因此，在每个集合中，动态主题模型依然按照 LDA 主题模型的建模过程进行。DTM 的图模型表示如图 4 所示。

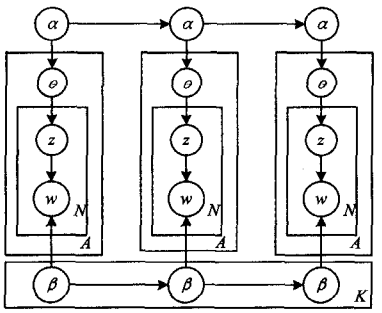

图 4　DTM 图模型

图 4 示出了对 3 个时间段的文本建模，将其与 LDA 模型的贝叶斯图模型相比可知，DTM 的贝叶斯图模型近似于按照时间先后顺序训练了 3 个独立的 LDA 模型，不同之处在于当前时间段内文本训练的主题模型的参数（α 和 β）受前一个时间段内文本的主题模型训练结果的影响。

DTM 将时间信息结合到主题模型中，得到的主题表现出随时间演化的趋势。DTM 的优点在于能够利用时间序列对整个文本集进行建模，从而揭示文本集中主题随时间的演化规律。然而，尽管 DTM 在主题动态演化方面取得了成功，但是在实际应用中还存在诸多不便，例如，DTM 没有考虑每个时间段内文档数目对主题数目的影响，也没有对时间段间主题的动态关系进行建模[24]；同时，DTM 存在如何寻找最优时间切片方式的问题。

3.2　扩展的时态主题模型

上文介绍的两种简单的时态主题模型都是在 LDA 模型的基础上进行扩展后得到的。虽然它们在动态文本流中都能取得较好的效果，但是依然不可避免地存在一些缺陷。目前，随着时态信息处理研究的发展，大批学者在简单时态主题模型的基础上进行了各种变形和拓展，以实现更好的时态应用。因此，本节将对这些模型的变化和拓展进行总结，以更好地反映时态主题模型在近几年的发展。

3.2.1　基于时间因素的改进

前面介绍的简单时态主题模型都是先根据时间信息按照时间顺序划分时间片，然后分别处理属于各个时间片内的文本集，最终根据各个时间片内的主题分布获取主题随时间的演化规律。为了解决时间片的离散性问题，研究者针对连续时间片因素对原有的时态主题模型进行了改进。

2007 年 Wei Xing 等人提出了动态混合模型 DMM（Dynamic Mixture Model）[25]，DMM 是基于条件概率的方法，对简单时态主题模型中同一时间段内文档的可交换属性进行了改进。与 DTM 和 TOT 相比，DMM 具有更强的时间假设，主要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档，并假设模型参数由前一时刻的混合分布生成。DMM 的演化依赖关系假设连续两篇文档中主题的分布存在演化关系，因此更适用于获取文本间更细微的内容和强度演化，但该模型对文档时间顺序有严格的限制，处理效率低。

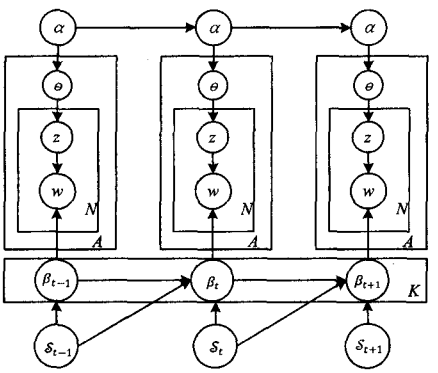

图 5　cDTM 图模型

2008 年 Wang 等人提出了一个持续时间动态主题模型 cDTM（continuous-time Dynamic Topic Model）[26]。图 5 描述了 cDTM 模型的生成过程。该模型使用布朗运动对主题的演化建模，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换为模型选择的问题，即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。同时，cDTM 采用 KalmanFilter 算法实现快速推断，从而优化了离散时间片中的内存消耗和模型计算复杂度。

3.2.2　基于非参数贝叶斯方法的变形

大部分基于非参数贝叶斯方法的变形主要是基于 Dirichlet Process[27] 的变形。基于 Dirichlet Process 的方法可以自动地学习出主题的数目[28]，因此可以在一定程度上解决主题模型中自动确定主题数目这个问题。

2008 年，Ahmed 等提出 TDPM（Temporal DirichletProcess Mixture Model）[29]，通过 Dirichlet Process 确定演化过程中每个时间窗中的主题个数。但是由于使用基于 Dirichlet Process 的方法，在模型构建过程中需要不断调整一些参数数值的设定，并且在实际中运行复杂度高。于是，2010 年，Ahmed 等又提出 iDTM（infinite Dynamic Topic Models）[30]，通过引入 HDP（Hierarchical Dirichlet Processes）[28] 方法，实现了对文本的潜在结构的建模，包括主题个数、主题分布以及主题趋势。iDTM 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；同时解决了单纯使用 LDA 过程中各时间窗内主题数固定的问题，并在多个领域得到了广泛的应用。iDTM 的图模型如图 6 所示。

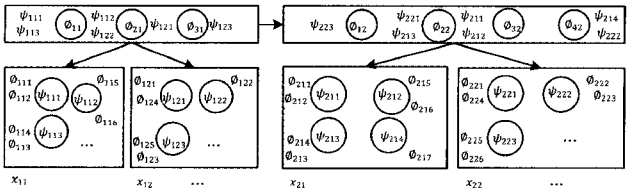

图 6 iDTM 图模型

3.2.3 基于有监督的学习方法

目前大部分主流的主题模型都是无监督的，只需要输入主题数目和文档集，模型就能进行主题的自动学习。因此，无监督学习方法得到的主题往往解释性较差，有时很难被理解[31]。为了解决无监督模型存在的一些问题，研究者们通过对无监督的文本生成模型进行适当的变形，得到有监督的主题模型。

2013 年 Elshamy 等人提出 ciDTM (continuous time infinite Dynamic Topic Models)[32]，通过构建 Dim Sum Process 主题生成过程，实现了在连续时间内可变主题数的主题演化模型，解决了在线实时媒体中需要根据文档的到达顺序而不能通过文档时间戳进行建模的问题，虽然 ciDTM 解决了多个问题，但是实现起来比较复杂。ciDTM 的图模型如图 7 所示。

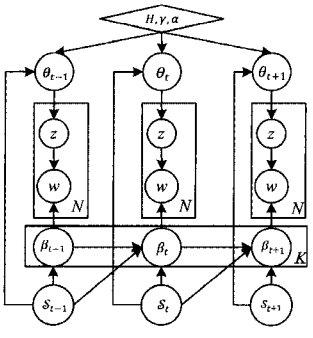

图 7 ciDTM 图模型

2015 年大连海事大学蒋卓人等人提出一种结合有监督学习的动态主题模型(Supervised Dynamic Topic Model, S-DTM)[33]。该模型不仅能够随时间的变化对语言进行动态建模，而且结合有监督学习技术，在主题变分推理中加入标签约束，从而建立主题与标签之间的映射关系，解决了主题难以表达解释和主题数目难以确定的问题，也使主题结构的可解释性大大增强。但在实际应用中，随着主题数目的增加，时间复杂度也随之增大。如何在保证结果精确性的前提下减少时间的消耗是接下来所需要解决的问题。

3.2.4 面向特定任务的改进

为了使时态主题模型满足特定的任务，研究者们在已有的研究基础上进行了改进，提出了面向特定任务的时态主题模型。

为了对基于文本流的在线主题演化模型进行研究，解决如何从海量文本中发现主题并进行演化分析，2008 年 AlSumait 等提出 OLDA(Online LDA)模型[34]，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，且利用演化矩阵实时地检测新主题的产生。但是 OLDA 因采用离散时间方式而导致适用领域有所限制。

2010 年国防科技大学的崔凯提出 iOLDA 模型[8]，它将 LDA 模型扩展到在线文本流上，将文档按时间片进行划分，用当前时间片的后验概率影响下一时间片的先验概率来保持在线分析中的主题连续性，利用增量 Gibbs 算法进行参数估计，获取连续的文档-主题和主题-词的概率分布。但由于模型构建的过程中使用的是迭代式推理算法，因此与处理要求的实时性还存在一定的差距。

2011 年 Nasir Naveed 等人提出 ATT (Author-Topic-Time model)[12]，通过融入用户和时间戳信息，对 LDA 主题模型进行扩展，不仅实现了对主题演化过程中主题趋势的分析，而且还能获取随着时间的变化用户信息的改变。该模型摆脱了 Author Topic Model[10] 只能适用于静态文档的局限性，通过引入时间因素，更加适用于论文集以及社交媒体中信息的检索与推荐，提高了模型的实用性。AT 和 ATT 的图模型分别如图 8 和图 9 所示。由图可以看出，AT 模型中每个作者对应于一个主题上的分布，并且所有的作者共享一个主题集合。而 ATT 模型同时结合文档的作者和时间戳信息，在 AT 模型的基础上增加了主题基于时间的分布。但由于该模型还未扩展到大量的文本集中，因此不能解决社交媒体等领域中潜在社区结构的演化。

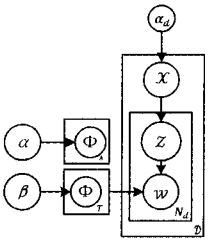

图 8 AT 图模型

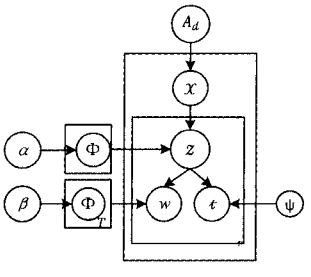

图 9 ATT 图模型

2013 年 Li Daifeng 等人提出了 Dynamic Community Topic Model(DCTM)[35]，该模型通过对 Community Topic Model 引入时间变量进行扩展，其主要通过伯努利分布表达用户主题在时间上的分布，最终不仅能够获取不同时间段中用户兴趣的改变，而且也能观察社区和主题基于时间的演化。

以上介绍的几种时态主题模型主要是针对已有的简单时态主题模型中所存在的问题，从各个不同的角度对其进行改进和扩展，每个模型不仅与具体的特定的任务有关，也与当前信息技术发展的程度有关。因此，随着研究者们对时态主题模型的不断研究，将来还会不断出现越来越多的改进的时态主题模型，表 2 是对上述介绍的扩展的时态主题模型的汇总，并对各个时态主题模型的优缺点进行了简单的描述。

表 2　扩展的时态主题模型汇总

扩展		时间	作者	模型	扩展模式	简单描述	优点	缺点
简单模型	基本时态的主题模型	2006	Wang 等[17]	TOT	LDA→TOT	主题随时间变化	能分析数据随着时间的结构变化	不具备在线处理能力
		2006	Blei 等[7]	DTM	LDA→DTM	主题会随着时间变化，且满足一阶马尔科夫假设	能够揭示主题随时间的演化规律	没有对时间段间主题的动态关系建模
	基于时间因素的改进	2007	Wei Xing 等[25]	DMM	TOT & DTM→DMM	主要针对多维时间序列的在线文本流	考虑动态文本流中数据的时间戳	对文档时间顺序有严格的限制，处理效率较低
		2008	Wang 等[26]	cDTM	DTM→cDTM	使用布朗运动模型实现主题的演化过程	实现了词和主题的分布演化	主题数固定
扩展的时态主题模型	基于非参数贝叶斯方法的改进	2008	Ahmed 等[29]	TDPM	LDA→TDPM	通过 Dirichlet Process 确定时间窗中的主题个数	主动学习主题数目	运行复杂度高
		2010	Ahmed 等[30]	iDTM	DTM→iDTM	引入 HDP 方法，通过数据推断，对文本的潜在结构建模	允许主题数持续变化	不适用主题变化较快的领域
	基于有监督的学习方法	2013	Elshamy[32]	ciDTM	iDTM→ciDTM	将 oHDP 与 cDTM 结合，在 HDP 中使用 Wiener process 对动态主题建模	从时间连续性、支持在线处理等方面实现主题演化	实现比较复杂
		2015	蒋卓人等[33]	S-DTM	DTM&Labeled-LDA→ S-DTM	应用于多标签时序语料库上的一个时间序列主题模型	解决了主题难以解释和主题数目难以确定的问题	时间复杂度高
	面向特定任务的改进	2008	AlSumait 等[34]	OLDA	LDA→OLDA	根据不断到来的新文档，增量构建新模型	能够实时地检测新话题的产生	适用领域有所限制
		2010	崔凯[8]	iOLDA	LDA & OLDA→iOLDA	将 LDA 模型扩展到在线文本流上	及时发现热点话题并预测其后续演化趋势	对实时性的要求还存在一定的差距
		2011	Nasir Naveed 等[12]	ATT	LDA & AT→ATT	结合文档的作者和时间信息，对 LDA 进行扩展	实现主题基于时间的演化分析	面向特定的任务，局限性大
		2013	Li Daifeng[35]	DCTM	CTM→DCTM	通过伯努利分布表达用户主题在时间上的分布	能同时观察社区和主题的演化	精确性低

4　基于时态主题模型的应用

由于时态主题模型几乎被用到了所有的文本挖掘和信息处理领域，其中受关注比较多的包括社交媒体分析、学术文献挖掘、社区发现等应用领域。本节将着重介绍时态主题模型在这几个领域的应用。

4.1　社交媒体中的应用

随着网络的快速发展，社交媒体也在飞速发展。以新浪微博为例，作为一种新型的媒体数据，其与传统的文档集合有着显著的不同。其由于更新速度快，因此带有明显的时间特征。如何根据时间因素从主题模型的角度来考虑具有时态信息的文本流的检索问题，是当前急需解决的问题。

4.1.1　时态主题模型在社交媒体中的发展

社交媒体信息具有极强的实时性。例如，用户在微博上阅读了一篇关于日本海啸的文章，若要继续查看微博中更多的关于此方面的信息，传统的基于关键词的检索没有聚焦于事件的热点，因此不适用于此类检索。而目前在信息检索领域分析主题演化中主题的趋势和用户的角色是一个很重要的挑战。

早期对社交媒体数据的研究没有内容，仅仅分析博客领域的图结构。Kumar 等通过超链接构建博主之间的联系图进行对博客领域建模，并根据图的属性分析图的演化[36]。随后提出的 Twitter-LDA[19]实际上是基于 AT 模型[10]的变化，它根据数据内容同时对用户层面的背景、Twitter 层面的背景进行建模[37]。但是，由于 Twitter 的长度过短，直接通过无监督学习很难得到理想的 Twitter 主题分布。最重要的是它忽略了社交媒体中的数据和时态信息是紧密联系的，它们假设这些维度是相互独立的，没有对文本的内容和时态信息共同建模，因此无法分析随着时间变化的主题的演化过程。

Avinava Dubey 等人通过扩展 TOT 模型，结合词和时间戳的非参数分布，提出了 npTOT 模型[38]。npTOT 模型解决了之前一些模型存在的两个问题：主题的概率分布的不灵活性和固定的主题数。npTOT 通过替代层级 Dirichlet 过程，避免了对主题数目的限制。它给予主题的非界限数一个后验概率，并确保文档间的主题是共享的，最终实现了在事先并不知道文本的主题数目的情况下利用 npTOT 模型获取文本集中基于时间而改变的主题。此外，npTOT 的一个应用优点是其可以很容易地扩展到高维协变量值，因此能对主题流行度中的地理位置变化建模。

ciDTM[32]对文本的主题数目的演化反映了在连续时间内文本集中的可变主题数，同时也增加了文本内容的丰富性，替代了之前将发现的词吸收到预先设定的主题中的方法，因此，处理的文本越多，最终发现的主题也会不断增多。

但是，在社交媒体中，由用户产生的内容通常是时态主题（例如突发事件）和稳定主题（例如用户兴趣）的混合，由于它们之间的不同特性，从社交媒体中的稳定主题内区分时态主题是非常重要并且非常有用的，上述所提到的几种模型虽然能够发现主题的演化，但不能直接用于从稳定主题中区分时态主题。2013 年北京大学的 Yin Hongzhi 等人提出的 UTM (User-topic mixture Topic Model)[39] 解决了这一问题。UTM 模型将用户和时态特征结合到一个混合模型中，能够直接用于同时监测稳定主题和时态主题。同时，与已有的时态主题监测方法的不同之处在于，UTM 能够自动抽取与时态特征相关的文本并将其聚簇到时态主题中，最终使用户能够根据实时需求检索到与此相关的内容。

4.1.2 时态主题模型在社交媒体中的应用挑战

在社交媒体中使用主题模型所面临的挑战主要来源于社交媒体中数据本身的特点所带来的挑战。

1)庞大的数据规模及文本长度过短。由于数据的超大规模,数据的收集工作也随之变得繁重,并且很难保证获得的数据是完整的;再加上文本长度的限制,以至于无法实现精确分析,最终可能会导致检索的效果不尽如人意。

2)更新速度快。迅速的更新速度是在线媒体的一个显著特征,但这也加大了对实时数据的获取难度。

3)噪音大。与普通的网络文本数据相比,在线社交媒体中用户往往书写随意,因此会产生更多的新生词、错别字、符号语言等。由于这些词的高频率出现,其对信息检索的任务造成了一定的影响[31]。

4.2 学术文献中的应用

主题模型在学术文献中的应用已经受到越来越多的关注。通过对学术文献的应用研究可以进一步理解学术界的发展及演化,同时对于了解科技的进步以及未来科学领域的发展趋势也是非常有意义的。

4.2.1 对学术引用的建模

主题演化的主要任务是对每一个时间段进行学习以得出其所对应的主题空间,同时发现相邻时间段主题的变化,以及同一个主题在不同的时间点内的变化。利用学术文献分析主题的演化主要是通过论文之间相互引用的链接。Qi He 等人讨论了对学术文献进行主题演化建模的最主要的 3 种思路[40]:第一种是时间独立的主题演化模式,即每个时间段内的主题仅与当前时间段内部的文献集有关。第二种是时间变化的主题演化模式,即每个时间点的主题不仅与当前的时间段内部的文献集有关,还与在此之前的时间段的所有文献集有关。第三就是引用限制的主题演化模式,即每个时间点的主题仅与当前的时间点的文章以及在此之前被这些文章引用的所有文章集合有关。

科学领域中主题演化表明了对某个主题的研究是如何相互影响的。因此,分析主题的逐渐演变的过程对理解学术研究是非常重要的。例如,主题演化分析有助于理解并客观地评价某位科学家或某篇论文对该研究领域的贡献。主题演化分析因在信息检索领域的重要性和极大的应用潜力最近几年受到了极大的关注。

学术文献不同于一般的新闻报道和博客文章,学术论文一般都需要附带非常详细的参考文献,且这些参考文献通常都与原来的论文有着非常紧密的联系。最近,利用主题模型观察文献之间的链接关系的研究越来越多。Zhou 等人直接利用 LDA 模型,通过观察学术文献中论文与论文的联系来分析时态主题的演化[41];同时,结合论文的内容维度,考虑论文之间的相似性与多样性,解释了为什么某些主题的趋势逐渐变弱而另一些主题则逐渐扩张。前面提到的 DTM 模型认为主题随着时间的改变逐渐演化并且呈正态分布。随后,Bi Chen 等人提出 c-ITM (citation Inheritance Topic Model)用于分析学术文献间的相互引用中的主题演化[42],同时考虑了文献中内容与文献间的引用,并能明确地抽取出时间变化与主题演化之间的关系。但与 DTM 模型不同的是,DTM 认为在时间点 t 时的论文都与时间点 $t-1$ 时的论文有关,而 c-ITM 模型则认为在时间点 $t-1$ 时被引用的论文仅与时间点 t 时这些论文所引用的论文有关。

4.2.2 对学术文献作者的建模

虽然上述提到的各种主题模型能够在主题演化方面产生良好的分析效果,但它们都忽略了文章的作者信息,实际上就是把所有的作者都看作是完全等同的,因此导致了应用中的限制。

早期提出的 AT 模型[10]是对 LDA 模型的一种扩展,它使用了一个基于主题的表现法对文档的主题和作者的兴趣进行建模。尽管 AT 模型在学术文献中的应用较之前的研究产生了巨大的影响,但其却忽略了学术文献中的时间维度,即每个主题随着时间的强度如何变化。为了能够同时利用学术文献中的时间戳信息和数据内容以及作者信息,Nasir Naveed 等人提出的 ATT 模型[12]实现了主题演化的分析。ATT 模型通过对 LDA 主题模型的扩展,结合文献中所包含的作者和时间戳信息,根据潜在的主题演化,以获取基于时间的改变和作者研究兴趣的变化。此外,ATT 模型在应用中的另一个优点是根据主题演化的信息,发现在主题演化的不同阶段中对该主题具有贡献的关键研究者。在实际应用中,ATT 模型可作为信息检索的有力工具。例如在社交检索和推荐系统中,可根据用户的研究兴趣,通过主题分析所获取的结果向用户推荐与其研究领域相关的研究者和参考文献。

4.3 数据社区中的应用

运用社区发现方法有效地发现数据中的社区结构及其演化过程有助于研究者更好地分析网络结构性质,了解整个网络的动态趋势,从而为实现网络结构的优化、资源的搜索和推荐等提供服务。在大规模网络中,社区发现的工作主要包括两部分:发现各个时间段的社区结构和通过社区间的关联找出社区的演化过程[43]。

网络数据由实体及描述实体属性之间的链接组成,如前面所提到的文献中的引用关系、网页中的链接关系等。然而目前大部分的研究主要聚焦于社区的结构属性,忽略了社区的主题特征和时态性。例如,某位研究者在某个阶段有两个不同的研究领域(语义网和信息检索),这意味着该研究者属于不同主题的两个社区。相反,根据社区的演化及主题,可以推断在不同的阶段该研究者的研究领域的变化。

为了能够发现社区和主题之间的动态关系,需要监测它们基于时间的变化情况。传统的方法将不同的时间戳看作是相互独立的,并忽略了连续时间戳之间的时态连续性。这样的研究方法存在两个问题:1)如何确定不同时间戳中潜在变量之间的对应关系。例如,对于在时间点 t 时刻的某个社区,我们很难知道该社区是否由上一个时间点 $t-1$ 演化而来;2)没有考虑持续时间戳之间的时态关联,例如某位研究者以前的研究领域是否影响当前的研究兴趣。

为了解决这两个问题,Li Daifeng 等人提出了 DCTM[35]。DCTM 根据连续时间戳之间的时态连续性来获取社区和主题的动态特征。DCTM 的基本思想是在对社区的演化建模中认为社区的分布是基于贝努利试验的,即当值为 1 时,认为当前的社区是由上个时间段的状态决定;当值为 0 时,认为当

前的社区不是由上个时间段决定的，而是一个新生的社区状态。最终实验表明，通过 DCTM 能发现具有相似主题的社区以及社区与主题的共同演化，从而能够更好地理解社区网络的动态性，并有助于个性化的推荐。但由于采用贝努利试验确定社区的演变状态使得每一时间段的社区演变具有随机性，最终会导致结果分析不精确。

Juan Bi 等人提出将动态主题模型用于社交网络中的社区发现，有助于理解社交网络中的动态特征[44]。与之前的方法不同，该方法将社区和主题看作不同的潜在变量，因此该模型不仅能够同时发现社区和主题，也能追踪所发现的社区和主题基于时间的演化。

上文总结了时态主题模型在具体领域中的应用，虽然时态主题模型已经被广泛地应用于上述领域中，但仍然存在一些受限制的缺点。将时态主题模型的分析结果以可视化形式呈现，会促进时态主题模型分析的可理解性，并推进时态主题模型的进一步应用。

5　发展趋势

尽管主题模型自提出以来已经发展了十余年，但是由于新需求的出现以及基础技术的革新，它仍然是一个非常有意义与有吸引力的研究方向。从第 3 节对时态主题模型的介绍中可以看出，近几年来随着学者们的不断研究与探索，对时态主题模型的研究工作集中在模型的改进和优化方面。除此之外，从第 4 节对时态主题模型的应用总结中可以看出，时态主题模型已经不局限于理论研究的阶段，近几年来时态主题模型开始逐渐应用于信息处理的各个领域，包括文本挖掘、信息检索等。另一方面，时态主题模型的研究仍然是一个相对不完善的研究领域，目前国内在该方面的研究仍然比较有限，研究时态主题模型的文献主要关注于早期提出的简单时态主题模型，因此时态主题模型未来的发展趋势应主要包括以下 3 个方面：时态主题模型性能的优化、时态主题模型优劣度的评价指标以及时态主题模型在信息处理方面的进一步应用。

5.1　时态主题模型性能优化

在时态主题模型性能优化方面，需要更高效的训练算法。目前大部分时态主题模型都是将词项空间变换到主题空间，区别在于主题模型表示上的差别，或者说是在最优化时使用的目标函数不同。由于通常无法求得精确解，参数估计问题至关重要，有多种算法可用于估算主题模型中的参数，常用的方法包括利用 EM 算法迭代计算贝叶斯后验分布的概率的变分推断方法[45]和利用马尔科夫链蒙特卡洛方法的 Gibbs 抽样算法[46]。Mimno 等人[47]提出一种基于任意图模型先验的吉布斯采样算法，它能够更好地处理大量文本集合训练推理过程中的复杂关联。Nallapati 等人[48]提出并行的变分 EM 算法来加速训练过程，该算法可用于主题模型中参数的求解。这些方法对大部分的主题模型具有通用性，因此对时态主题模型也适用。但随着时态信息的广泛应用，如何设计一种具体针对时态主题模型的性能优化方法还需要进一步的研究。

5.2　时态主题模型优劣度的评价指标

如何客观地衡量主题模型的优劣度已经成为主题模型领域所面临的一个主要问题。对主题模型的评估一直以来都受关注，但到目前为止依旧还未得到很好的解决，主要原因是主题模型具有非监督的特性，不同的应用程序中需要实现不同的任务，例如信息检索、文档分类等，使得模型的选择变得困难，因此很难直接评估一个模型的表示方法的好坏[31]。早期的一些评估方法包括调和平均数法、文档实现方法等，但这些模型基本上是不精确的[49]。因此目前用于估计主题模型效果的标准主要是复杂度比较，大多数情况下研究人员通过使用复杂度对模型进行定量评估[50]。即若新的模型的复杂度小于已存在的模型，就认为该模型的建模效果更好。然而基于复杂度的评估方式存在的一个最大的问题就是一个复杂度较低的模型未必可以很好地得到主题词。因此，2009 年 Hanna M. Wallach 等人在比较已有的一些评估主题模型的方法后，提出了包含精确度和有效性在内的两种可选方法。实验结果表明，与已经存在的模型相比，新提出的模型具有更好的精确性[49]。同时，Wray Buntine 在 Hanna M. Wallach 等人的基础上，通过转换 Wallach 的 left-to-right 算法，提出了新的估计文档集合的似然方法，该方法更加公平[51]。除了这些方法以外，研究人员提出可以利用一些任务实现间接评估模型，包括对文档相似度的计算、主题间相似度的计算等[31]。但这种方法仍然存在很多问题，由于每个时态主题模型的应用领域不同，时间片的划分不同，因此很难判断出模型之间的好坏。总而言之，近几年来对时态主题模型效果的评估方法的研究依旧存在很多不足，如何改进现有的或提出新的时态主题模型优劣度评估方法将成为未来的重要研究领域。

5.3　时态主题模型在信息处理领域的应用

时态主题模型在信息处理领域的进一步应用将是另一个发展趋势。时态主题模型本质上是一种对具有时间信息的文本概率建模的方法，因此可以应用在信息处理领域的各个方面。Freddy Chong Tat Chua 等人[52]提出 LDTM (Linear Dynamical Topic Model)，考虑到用户与其好友之间话题的相关度较高，在主题训练时考虑好友之间的兴趣影响，通过主题抽取获取用户兴趣随时间变化的分布，最后在此预测基础上实现更有针对性、更加全面高效的社会化推荐系统。徐桂彬等人[53]通过改进相关主题模型(CTM)使其具有动态性，提出了动态相关主题模型 DCTM (Dynamic Correlated Topic Model)，将 DCTM 作为降维模型与隐马尔克夫模型相结合对音乐分类。由于 DCTM 的动态建模更好地提取了对分类有用的信息，因此增强了该方法的分类能力，最终建立了真实的满足人们实际需要的音乐信息检索系统。袁柳等人[54]基于 LDA 主题模型提出了 Temporal LDA 主题模型，其借助标签数据的时态特征以及基于时态特征的标签间语义关联进行分析，并提出发现标签时态特征的时间段划分准则，最终将其用于分析数据时态特征对所生成主题的影响以及标签预测。尽管近几年对标签学习问题的研究兴起了一股热潮，但相关研究中标签的时态特征却很少受到关注。因此，时态主题模型在不同领域的高效应用将成为当前及未来值得深入研究的一大热点。

总之，在使用时态主题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，并选取合适的模型评估方法对模型的优劣度进行评估，通过不断地调整时间片设置、内容粒度等参数使得模型达到结果最优。同时，在当前模型不能满足需求的情况下，需要重新考虑模型的改进和完善。

结束语 随着网络的不断发展，数据大多以动态文本流的方式出现，因此时态主题的发现及演化分析吸引了越来越多研究人员的关注。本文对时态主题模型及其应用进行了综述，首先介绍了主题模型和时态主题模型的基本知识，并对近年来各种时态主题模型进行了分类，最后总结了时态主题模型的应用并对未来的发展趋势进行了展望。

参 考 文 献

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

[2] CHEN B. Topic Oriented Evolution and Sentiment Analysis [D]. The Pennsylvania State University, 2011.

[3] 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.

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

[5] BLEI D M, LAFFERTY J D. Topic models In Text Mining: Classification Clustering and Applications [M]. Chapman & Hall, London, UK, 2009: 71-94.

[6] WANG L. The Research of Dynamic Network Community Detection Algorithm [D]. Shenyang: Northeastern University. 2013. (in Chinese)
王玲. 动态网络社区发现算法研究[D]. 沈阳：东北大学，2013.

[7] BLEI D M, LAFFERTY J D. Dynamic Topic Models [C]// Proceedings of the 23rd International Conference on Machine Learning. ACM, 2006: 113-120.

[8] CUI K. The Research and Implementation of Topic Evolution Based on LDA [D]. Changsha: National University of Defense Technology, 2010. (in Chinese)
崔凯. 基于 LDA 的主题演化研究与实现[D]. 长沙：国防科学技术大学，2010.

[9] XU G, WANG H F. The Development of Topic Models in Natural Language Processing [J]. Chinese Journal of Computers, 2011, 34(8): 1423-1436. (in Chinese)
徐戈，王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报，2011，34(8)：1423-1436.

[10] ROSEN-ZVI M, GRIFFITHS T, STEYVERS M, et al. The Author Topic Model for Authors and Documents [C]// Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2004.

[11] BLEI D M, LAFFERTY J D. Correlated Topic Models [C]// Proceedings of the 23rd International Conference on Machine Learning. 2006.

[12] NAVEED N, SIZOV S, STAAB S. ATT: Analyzing Temporal Dynamics of Topics and Authors in Social Media[C]// Proceedings of the 3rd International Web Science Conference. ACM, 2011.

[13] BLEI D, MCAULIFFE J. Supervised topic models[J]. Advances in Neural Information Processing Systems, 2010, 3: 327-332.

[14] MCCALLUM A, CORRADA-EMMANUEL A, WANG X. The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email [OL]. http://ciir-publications.cs.umass.edu/pdf/IR-381.pdf.

[15] NALLAPTI R, COHEN W. Link-PLSA-LDA: A New Unsupervised Model for Topics and Influence in Blogs[C]// Proceedings of the International Conference for Weblogs and Social Media. Seattle, Washington, USA, 2008.

[16] RAMAGE D, HALL D, NALLAPATI R, et al. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora[C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore, 2009: 248-256.

[17] WANG X R, 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. New York, NY, USA, ACM, 2006: 424-433.

[18] ALSUMAIT L, BARBARÁ D, DOMENICONI C. On-Line LDA: Adaptive Topic Models for Mining Text Streams withApplications to Topic Detection and Tracking [C]// Proceedings of the 8th IEEE International Conference on Data Mining. IEEE, 2008: 3-12.

[19] ZHAO W X, JIANG J, WENG J, et al. Comparing Twitter and Traditional Media Using Topic Models[M]// Advances in Information Retrieval. Springer Berlin Heidelberg, 2011: 338-349.

[20] CHANG J, BLEI D M. Relational Topic Models For Document Networks[C]// AISTATS. 2009.

[21] LI Dai-feng, DING Ying, XIN Shuai, et al. Adding Community and Dynamic to Topic Models[J]. Journal of Informetrics, 2012, 6(2): 237-253.

[22] LIAO J H, SUN K Y, ZHONG L X. Study on a Hot Topics Analysis System based on Time Sliced Topic Model[J]. Library and Information Service, 2013, 57(9): 96-102. (in Chinese)
廖君华，孙克迎，钟丽霞. 一种基于时序主题模型的网络热点话题演化分析系统[J]. 图书情报工作，2013，57(9)：96-102.

[23] DING W, CHEN C. Dynamic Topic Detection and Tracking: A Comparison of HDP, C-word, and Co-citation Methods[J]. Journal of the Association for Information Science and Technology, 2015, 65(10): 2084-2097.

[24] FAN Y M, MA J X. Review on the LDA-based Techniques Detection for the Field Emerging Topic[J]. New Technology of Library and Information Service, 2012(12): 58-65. (in Chinese)
范云满，马建霞. 利用 LDA 的领域新兴主题探测技术综述[J]. 现代图书情报技术，2012(12)：58-65.

[25] WEI X, SUN J, WANG X. Dynamic Mixture Models for Multiple Time-Series [C]// Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, India,

2007：2909-2914.

[26] WANG C，BLEI D，HECKERMAN D. Continuous Time Dynamic Topic Models [M]// D Mallester A Nicholson，Uncertainty in Artificial Intelligence. 2012：579-586.

[27] TEH Y W. Dirichlet processes[M]// Encyclopedia of Machine Learning. Springer，2010.

[28] TEH Y W，JORDAN M I，BEAL M J，et al. Hierachical Dirichlet process[J]. Journal of the American Statiscal Association，2006，101(476)：1566-1581.

[29] AHMED A，XING E P. Dynamic Non-Parametric Mixture Models and the Recurrent Chinese Restaurant Process：With Applications to Evolutionary Clustering[C]// Proceedings of the SIAM International Conference on Data Mining. Atlanta，Georgia，USA，2008：219-230.

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

[31] ZHAO X，LI X M. The Application of Text Mining Based on Topic Model[D]. Beijing：Peking University，2011. (in Chinese)
赵鑫，李晓明. 主题模型在文本挖掘中的应用[D]. 北京：北京大学，2011.

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

[33] JIANG Z R，CHEN Y，GAO L C，et al. A Supervised Dynamic Topic Mode[J]. Acta Scientiarum Naturalium University Pekinensis，2015，51(2)：367-376. (in Chinese)
蒋卓人，陈燕，高良才，等. 一种结合有监督学习的动态主题模型[J]. 北京大学学报，2015，51(2)：367-376.

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

[35] LI D F，YING D，XIN S，et al. Adding Community and Dynamic to Topic Models[J]. Journal of Informetrics，2012，6(2)：237-253.

[36] KUMAR R，RAGHAVAN N P，TOMKINS A. On the Bursty Evolution of Blogspace[C]// Proceedings of the International Conference on World Wide Web. 2003.

[37] ZHAO W X，JIANG J，WENG J S，et al. Comparing twitter and traditional media using topic models[C]// ECIR. 2011：338-349.

[38] DUBEY A，HEFNY A，WILLIAMSON S，et al. A non-parametric mixture model for topic modeling over time (2012)[J]. arXiv preprint arXiv：1208. 4411.

[39] YIN H Z，CUI B，LU H，et al. A Unified Model for Stable and Temporal Topic Detection from Social Media Data[C]// IEEE International Conference on Data Engineering. 2013：661-672.

[40] HE Q，CHEN B，PEI J，et al. Detecting topic evolution in scientific literature：how can citations help? [C]// Proceeding of the 18th ACM Conference on Information Andknowledge Management(CIKM'09). New York，NY，USA，2009：957-966.

[41] ZHOU D，JI X，ZHA H，et al. Topic Evolution and Social Interactions：How Authors Effect Research[C]// Proceedings of the 15th International Conference on Information and Knowledge Management. 2006：248-257

[42] CHEN B. Topic oriented evolution and sentiment analysis[D]. PA，USA：Tennsylvania State University，2011.

[43] YAN J. Research on Community Discovery Based on Topic Model[D]. Chengdu：Southwest University，2012. (in Chinese)
严姣. 基于主题模型的社区发现研究[D]. 成都：西南大学，2012.

[44] 毕娟，秦志光，黄嘉. Dynamic Topic Model for Detecting Community in Social Networks[C]// 全国博士生学术年会. 2013.

[45] WAINWRIGHT M J，JORDAN M I. Graphical Models，Exponential Families，and Variational Inference[J]. Foundationa and Trends in Machine Learning，2008，1(1/2)：1-305.

[46] GRIFFITHS T. Gibbs Sampling In the Generative Model of Latent Dirichlet Allocation [OL]. http：//www-psych. stanfordedu/～gruffydd/cogsci02/1.

[47] MIMNO D，WALLACH H，MACLLUM A. Gibbs sampling for logistic normal topic models with graph-based priors[C]// Proceedings of the NIPS Workshop on Analyzing Graphs. Whistler，Canada，2008.

[48] NALLAPATI R，COHEN W，LAFFERTY J. Parallelized variational EM for latent dirichlet allocation：An experimental evaluation of speed and scalability[C]// Proceedings of the ICDM Workshop on High Performance Data Mining. Omaha，USA，2007：349-354.

[49] WALLACH H M，MURRAY I，SALAKHUTDINOV R，et al. Evaluation methods for topic models[C]// ICML. 2009：1105-1112.

[50] HONG L，YIN D，GUO J，et al. Tracking trends：incorporating term volume into temporal topic models[C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM，2013：484-492.

[51] BUNTINE W. Estimating Likelihoods for Topic Models[M]// Advances in Machine Learning. Springer Berlin Heidelberg，2009：51-64.

[52] CHUA F C T，OENTARYO R J，LIM E P. Using Linear Dynamical Topic Model for Inferring Temporal Social Correlation in Latent Space[J]. Computer Science，2015，36(19)：189-221

[53] XU G B，DENG W. Music classification method combining DCTM and HMM[J]. Computer Engineering and Design，2012，33(11)：4245-4249. (in Chinese)
徐桂彬，邓伟. 结合 DCTM 与 HMM 的音乐分类方法[J]. 计算机工程与设计，2012，33(11)：4245-4249.

[54] YUAN L，ZHANG L B. Applying Temporal Features of Social Tags to Tag Predication[J]. Computer Science，2012，39(6)：179-183. (in Chinese)
袁柳，张龙波. 标签时态特征分析及其在标签预测中的应用[J]. 计算机科学，2012，39(6)：179-183.