概率主题模型综述

韩亚楠 刘建伟 罗雄麟
(中国石油大学(北京)自动化系 北京 102249)

摘 要 主题模型是当下文本挖掘中最主要的技术之一,广泛应用于数据挖掘、文本分类以及社区发现等.由于其出色的降维能力和灵活的易扩展性,成为自然语言处理领域的一个热门研究方向.Blei 等人提出了以 Latent Dirichlet Allocation(LDA)为代表的概率主题建模方法,在该模型中主题可以看作是单词的概率分布,主题模型通过单词项在文档级的共现信息提取出与文档语义相关的主题,实现将高维的单词空间映射到低维的主题空间,进而完成对目标文本数据的降维处理,开创了文本挖掘研究的新方向.其中 LDA 作为一种概率生成模型很容易被扩展为其它各种形式的模型,鉴于概率主题模型的应用价值、理论意义和未来的发展潜力,本文首先系统性地对 LDA 模型进行介绍,进而对基于 LDA 模型的各类扩展模型进行详细分类,并对其中各类的典型代表进行详细介绍,指出了各个概率主题模型被提出的原因以及其模型的具体形式、所具有的优缺点、适宜解决的问题等,进而又指出近年来主题模型典型应用场景;此外,本文还对目前概率主题模型常用的几个公认的数据集、评测方法以及典型实验结果进行详细介绍,并在最后指明了概率主题模型在进一步研究中需要解决的问题以及未来可能的发展方向.

关键词 主题模型;文本挖掘;LDA;高维数据;自然语言处理
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2021.01095

A Survey on Probabilistic Topic Model

HAN Ya-Nan LIU Jian-Wei LUO Xiong-Lin
(*Department of Automation, China University of Petroleum, Beijing* 102249)

Abstract Topic model is one of the most important techniques in text mining, which is widely used in data mining, text classification and community discovery. Topic model has become a hot direction in the field of natural language processing because of the excellent ability of the dimensionality reduction and the flexible ability to construct other probilistic models. Blei et al proposed LDA which is known as the most typical topic model. In this model, a topic is regarded as probabilistic distribution of words. Topic models extract semantic topics using co-occurrence of terms in document level, and are used to map high-dimensional word vectors to low-dimensional topic spaces, obtaining the low dimensional representation of documents. Topic models create a new direction of texting processing for data mining. As a probabilistic generative model, LDA can be easily extended to other models. Therefore, in view of the application value, theoretical significance and future development potential of probabilistic topic model, firstly, this paper systematically introduces the LDA model, making a particular categorization on topic models derived from LDA, and then points the motivation of every topic model, the advantages of every topic model, the problems that every topic model can solve, the form of every topic model, and the typical application scenarios that topic models can be used. In addition, several common datasets, evaluation metrics and typical experimental results of probability topic models are introduced

收稿日期:2020-02-26;在线发布日期:2021-02-26.**韩亚楠**,博士研究生,主要研究方向为机器学习.E-mail:857182813@qq.com.
刘建伟,博士,副教授,主要研究方向为机器学习、模式识别与智能系统、复杂系统的分析、预测与控制、算法分析与设计.**罗雄麟**,博士,教授,主要研究领域为智能控制、复杂系统分析、预测与控制.

in detail. Finally, we reveal the problems and the research directions of the probabilistic topic models in the future.

Keywords topic model; texting mining; Latent Dirichlet Allocation (LDA); high-dimensional data; natural language processing

1 引 言

随着时代的进步,以互联网为代表的信息技术已经获得飞速的发展,但是如何从互联网海量的文本数据中,例如在线网页、微博、新闻、学术文章等,快速、准确、全面地提取所需的信息是目前机器学习领域所面临的一大挑战.为了快速准确地在海量的文本中提取有用的信息,研究者希望能够找到一个词语或者一段话来对这些海量文本进行概括总结.

主题模型通过将高维单词空间映射到低维目标主题空间,有效地发现文档潜在的结构和隐藏的语义信息,最终实现对目标文档的降维处理、信息总结和摘要.其中,文本的降维表示能够使人们更好地理解文本的主要信息,使读者能够快速准确地理解文本集所讨论的主题内容.早期 Kontostathisa 等人提出了基于奇异值分解(Singular Value Decomposition, SVD)的潜在语义分析(Latent Semantic Analysis, LSA)[1]模型,在实现对文档的降维处理的同时,有效地实现对文档信息的总结提取.然而,该模型却面临"一词多义"和"多词一义"等问题.随后,为了解决 LSA 模型面临的问题,以及能够有效避免复杂的代数计算.Hofmann 提出概率潜在语义分析(Probabilistic Latent Semantic Analysis, PLSA)模型[2],在该模型中不考虑词序,文本语料可以由单词和文档的共现矩阵表示.从观测的单词中推断两个参数:一个是将语料库中文档联系起来的全局参数,代表了给定主题后单词出现的概率;另一个是每篇文档的参数,代表了文档的主题概率分布.PLSA 模型通过引入概率统计思想,大大降低了模型的计算成本.但是,在 PLSA 模型中,对特定文档中的主题的混合比例权重没有做任何假设.因此,在实际训练时常出现过拟合的情况.

Blei 等人在此研究基础上,提出了一个三层次的概率主题模型 LDA(Latent Dirichlet Allocation),在该模型中,通过一个概率生成模型把所有的文档参数联系起来,在 LDA 模型中,每篇文档中的主题的多项分布是通过一个与语料库中所有文档相关联的 Dirichlet 先验生成,实现模型的彻底"概率化"[3].自 2003 年 LDA 主题模型被提出后,主题模型才得到了广泛深入地研究,并在机器学习领域获得广泛应用.然而,由于现代计算机信息存储和处理性能的急剧增强,大型文本语料变得日益丰富,在实际应用中,传统的主题模型已经无法满足日益发展的需要.因此,一系列的新型模型相继提出.值得注意的是,当下对主题模型的研究主要包括基于 LDA 的概率主题模型、基于非 LDA 主题模型以及结合神经网络的深度学习主题模型.LDA 模型对主题模型的发展具有开创意义,所以,本文主要对基于 LDA 的各类扩展模型进行系统综述.例如,在线媒体文本具有明显的时间属性,往往需要结合时间信息对主题进行分析;在线评论具有明显的主观色彩,因此需要对评论的情感进行分析;在数据的社区发现中,文本内容和结构间具有明显的相关性,模型需要对文本进行相关性分析等.值得关注的是,概率主题模型因其良好的拓展性和广泛的应用前景使其在近年来已受到众多国外学者的青睐,基于 LDA 模型概率主题模型的分类图如图 1 所示.

本文第 1 节为引言;第 2 节为基本数学概念介绍;第 3 节介绍基于 LDA 的概率主题模型,其中首先介绍 LDA 主题模型,然后介绍基于 LDA 主题模型的概率扩展模型;第 4 节对近年来提出的几类典型的神经网络主题模型进行详细介绍;第 5 节对非 LDA 扩展模型进行简要介绍;第 6 节从文本分类、社交媒体、社区发现和图像处理等方面介绍概率主题模型的应用;第 7 节对目前概率主题模型常用的几个公认的数据集、评测方法以及典型实验结果进行详细介绍;第 8 节指出概率主题模型未来研究方向;第 9 节是本文的总结.

2 基本数学概念介绍

(1)伯努利分布(Bernoulli Distribution)

伯努利分布,也称为两点分布或者 0-1 分布,是

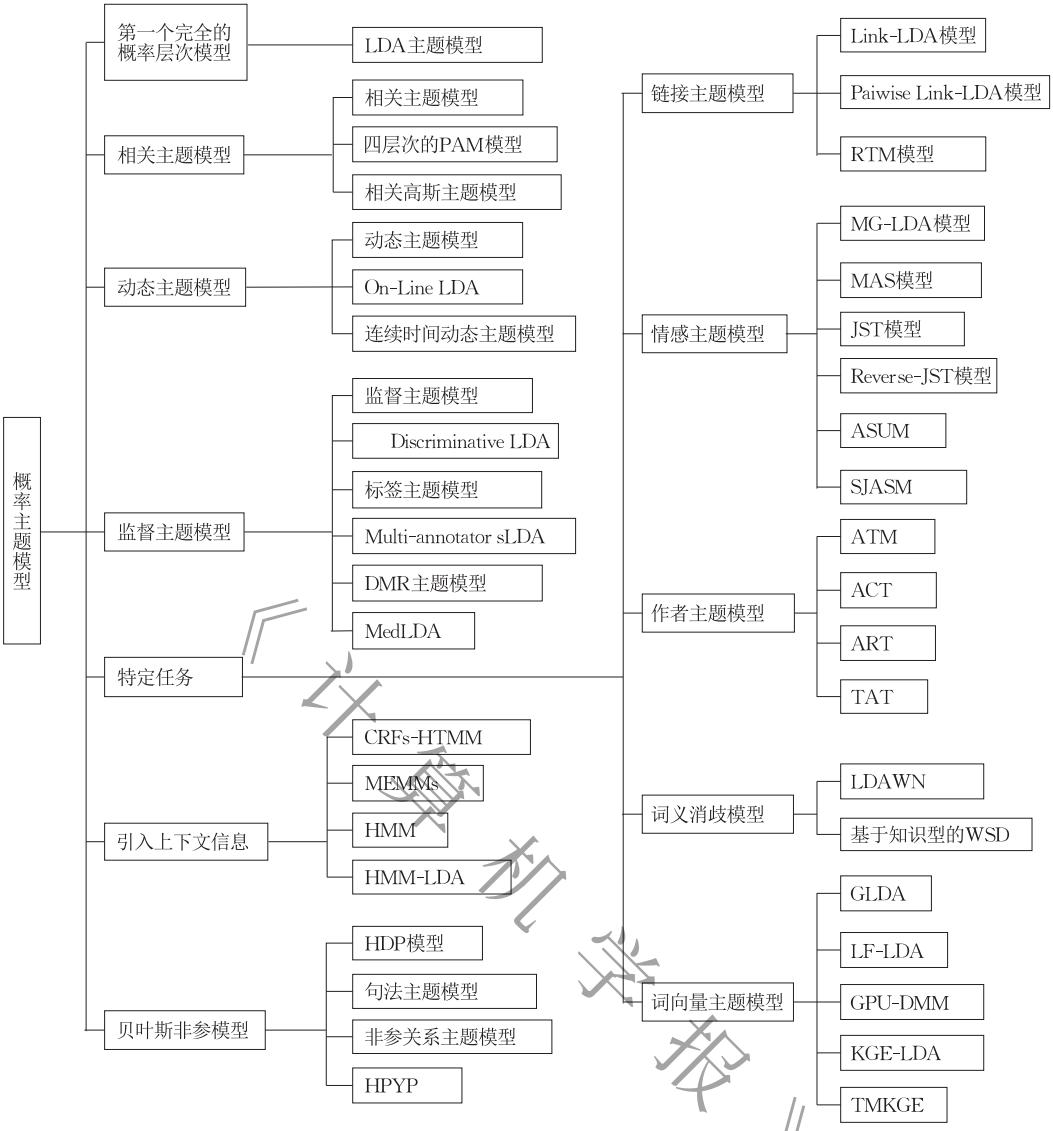

图 1　概率主题模型分类图

最简单的离散型概率分布，成功的概率记为 $p(0\leq p\leq 1)$，失败的概率记为 $1-p$.

(2) 二项分布(Binomial Distribution)

在概率论和数理统计学中，二项分布是 n 个独立成功/失败实验中实验成功的次数的离散概率分布，其中每次实验成功的概率为 p. 如果随机变量 X 服从参数为 n 和 p 的二项分布，记为 $X\sim B(n,p)$.

(3) 多项分布(Multinomial Distribution)

多项分布是二项分布的推广，如果一个随机变量 $X=(X_1,X_2,\cdots,X_n)$，满足下列条件：

① $X_i\geq 0(1\leq i\leq n)$，且 $X_1+X_2+\cdots+X_n=1$；

② $m_1,m_2,\cdots,m_n$ 为任意非负整数，且 $m_1+m_2+\cdots+m_n=N$.

那么，事件 $\{X_1=m_1,X_2=m_2,\cdots,X_n=m_n\}$ 的概率为

$$P\{X_1=m_1,X_2=m_2,\cdots,X_n=m_n\}=\frac{N!}{m_1!m_2!\cdots m_n!}p_1^{m_1}p_2^{m_2}\cdots p_n^{m_n} \tag{1}$$

其中，$p_i\geq 0(1\leq i\leq n)$，$p_1+p_2+\cdots+p_n=1$. 则称随机变量 $X=(X_1,X_2,\cdots,X_n)$ 服从多项分布.

(4) 高斯分布(Gaussian Distribution)

在概率论中，高斯(或正态)分布是实值随机变量(real-valued random variable)的一种连续概率分布，也是最为常见的分布之一，其概率密度函数的一般形式为

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{2}$$

其中，参数 μ 为分布的均值或期望；σ 为标准差.

(5) 逻辑斯蒂-正态分布(Logistic-Normal Distribution)

在概率论中，通过对多元正态分布进行逻辑斯

蒂变换,产生一个在 d 维单形(simplex)上的分布,称该分布为逻辑斯蒂-正态分布(Logistic-Normal Distribution).该分布的概率密度函数为

$$f_X(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{|2\pi\boldsymbol{\Sigma}|^{\frac{1}{2}}}\frac{1}{\prod_{i=1}^{D}x_i}\mathrm{e}^{-\frac{1}{2}\left\{\log\left(\frac{\boldsymbol{x}_{-D}}{x_D}\right)-\boldsymbol{\mu}\right\}^{\top}\boldsymbol{\Sigma}^{-1}\left\{\log\left(\frac{\boldsymbol{x}_{-D}}{x_D}\right)-\boldsymbol{\mu}\right\}},\boldsymbol{x}\in\mathcal{S}^D \tag{3}$$

其中,$\boldsymbol{x}_{-D}$表示 $\boldsymbol{x}$ 的第一个$(D-1)$分量向量;$\mathcal{S}^D$表示 D 维概率向量形成的单形;$\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 分别表示多元正态分布的期望和方差.

(6)狄利克雷分布(Dirichlet Distribution)

多项分布的共轭分布称为狄利克雷分布,也就是该分布的分布函数与多项分布分布函数具有相同的形式,在概率和统计学中,狄利克雷分布常表示为 Dir($\boldsymbol{\alpha}$).其概率密度函数如下式所示.

$$f(x_1,\cdots,x_K;\alpha_1,\cdots,\alpha_K)=\frac{1}{\boldsymbol{B}(\boldsymbol{\alpha})}\prod_{i=1}^{K}x_i^{\alpha_i-1}$$

$$\boldsymbol{B}(\boldsymbol{\alpha})=\frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)},\ \boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_K) \tag{4}$$

其中,参数 $\alpha_1,\cdots,\alpha_K>0$,并且对于所有 $i\in[1,K]$,满足$\sum_{i=1}^{K}x_i=1$ 和 $x_i\geqslant 0$.

(7)冯米塞斯分布(von Mises distribution)

在概率论和方向统计中,von Mises 分布,又称为圆正态分布或 Tikhonov 分布,是圆上的连续概率分布,是正态分布的圆形模拟.角度为 x 的 von Mises 分布的概率密度函数为

$$f_X(x\,|\,\mu,\kappa)=\frac{\mathrm{e}^{\kappa\cos(x-\mu)}}{2\pi I_0(\kappa)} \tag{5}$$

其中,$I_0(\kappa)$为 0 阶 Bessel 修正函数;μ 是位置的度量,表示分布集中在 μ 的周围程度;κ 是一种浓度的度量,$1/\kappa$ 与高斯分布中的 σ^2 作用相似.

(8)softmax 函数

softmax 函数,也称为指数归一化函数,它是一种 Logistic 函数的归一化形式,可以将 k 维实数向量压缩成区间[0,1]的新的 K 维实数向量,其函数形式为

$$\sigma(z)_j=\frac{\mathrm{e}^{z_j}}{\sum_{k=1}^{K}\mathrm{e}^{z_k}},\ j=1,\cdots,K \tag{6}$$

(9)sigmoid 函数

sigmoid 函数,也称为 Logistic 函数,常作为隐层神经元的激活函数使用,取值范围为(0,1),它可以将一个实数映射到(0,1)的区间,因此,可以用来构造二分类,其函数公式定义形式为

$$S(x)=\frac{1}{1+\mathrm{e}^{-x}} \tag{7}$$

(10)KL 散度(Kullback-Leibler divergence)

KL 散度,又称为相对熵或信息度量,是两个概率分布间差异的非对称性度量函数.典型情况下,P 表示样本真实分布,Q 表示样本的近似分布,则 KL 散度表示为

$$D_{\mathrm{KL}}(P\|Q)=\sum_{x\in X}P(x)\log\frac{P(x)}{Q(x)}.$$

(11)Wasserstein 距离

Wasserstein 距离是两个概率分布之间的距离度量函数,定义为

$$W(P_1,P_2)=\inf_{\gamma\sim\prod(P_1,P_2)}E_{(x,y)\sim\gamma}[\|x-y\|],$$

其中,$\prod(P_1,P_2)$是 P_1和 P_2联合分布的集合,对于每一个可能的联合分布 γ,从中采样得到样本 x 和 y,并计算出样本的距离$\|x-y\|$,进而可以计算在联合分布 γ 下的期望 $E_{(x,y)\sim\gamma}[\|x-y\|]$,在所有可能的联合分布中能够对这个期望值取到的下确界 $\inf_{\gamma\sim\prod(P_1,P_2)}E_{(x,y)\sim\gamma}[\|x-y\|]$就是 Wasserstein 距离.

(12)布朗运动

在概率统计中,称实随机过程 $W=\{W_t,t\geqslant 0\}$ 是标准的布朗运动,如果满足以下:

① $W_0=0$;

② 对于任意 $0\leqslant s<t$,$W_t-W_s\sim N(0,t-s)$;

③ W 具有独立增量性.

(13)Dirichlet 过程(Dirichlet Process,DP)

DP 主要参数为一个基分布 G_0和一个正的标量参数 α,常表示为 DP(α,G_0).实质上,DP 是分布上的分布,在该过程中,首先假设从一个 DP 随机抽取一个样本分布 G,之后,从 G 中独立地抽取 M 个随机变量$\{\Theta_m^*\}_{m=1}^{M}$,其表示形式如下所示:

$$\begin{aligned}&G\,|\,\alpha,G_0\sim\mathrm{DP}(\alpha,G_0)\\&\Theta_m^*\,|\,G\sim G,\ m=1,\cdots,M\end{aligned} \tag{8}$$

(14)Pitman-Yor 过程(Pitman-Yor Process,PYP)

PYP 与 DP 相似,作为 DP 的扩展,在聚类过程中能够直接从样本数据自动地识别样本类别个数,因此,模型无需事先给出样本个数.在该过程中,首先从 PYP 中随机抽取形成概率分布 G,之后,从 G 中独立地抽取 M 个随机变量$\{\Theta_m^*\}_{m=1}^{M}$,其表示形式如下所示:

$$G|\alpha,G_0\sim \mathrm{PYP}(d,\alpha,G_0)$$
$$\Theta_m^*|G\sim G,\ m=1,\cdots,M \tag{9}$$

其中，$d\in[1,0)$是 PYP 的折扣因子，$\alpha>-d$ 是强度参数，G_0为基概率分布.

(15) 中餐馆过程(Chinese Restaurant Process)

在概率论中，中餐馆过程是一个离散时间随机过程. 该随机过程是一个序列化的构造过程，随机过程中第 n 个时刻的值决定了如何对集合$\{1,2,3,\cdots,n\}$进行分组以及每组中包含的元素的个数，令 B_n表示第 n 个时刻分组的结果，B_n的概率分布是通过如下方式确定的. 初始时刻 $n=1$，以概率 1 获得其一般划分$\{\{1\}\}$. 在时刻 $n-1$，第 n 个新到的元素要么添加到 B_{n-1}中现有的某一分组中，要么新增加一个分组，最终形成分组 B_n，具体规则为

① 以概率$\frac{|b|}{n}$添加第 n 个新到的元素到 B_{n-1}的原有某一个分组中去，其中 $|b|$表示 B_{n-1}中分组的个数.

② 以概率$\frac{1}{n}$新增加一个分组，该分组只包含第 n 个新到的元素.

如此生成的随机过程具有可交换性，也就是说，重新对$\{1,2,3,\cdots,n\}$的顺序进行调整，不会改变 B_n的概率分布，而且从$\{1,\cdots,n\}$中删除元素 n 而获得的 $n-1$ 个元素形成的分布与 B_{n-1}是一致的. 时刻 n 的随机分组与时刻 $n-1$ 的随机分组的规律相同. 分配给任何特定分组的概率是：

$$P_r(B_n=B)=\frac{\prod_{b\in B}(\#b-1)!}{n!} \tag{10}$$

其中 $b\in B$ 是分组索引，$\#b$ 表示第 $b\in B$ 组中元素的个数.

3　概率主题模型

3.1　LDA 主题模型

Blei 等人在 2003 年提出了一个全贝叶斯的概率主题模型 LDA(Latent Dirichlet Allocation)[3]，它是一种用于离散数据集合(如文本语料库)的生成式概率模型. LDA 是一个三个层次的贝叶斯概率主题模型①，把每一个数据集都认为是一组潜在主题的混合，在文本建模过程中，利用主题概率分布对每篇文档进行摘要表达. 此外，LDA 模型基于“词袋”模型(bag of words)，词袋在建模过程中忽略单词的词序，进而对问题进行了简化处理. “词袋”假设不仅简化了模型复杂度，同时也为模型的进一步改进提供了契机. LDA 图模型如图 2 所示，其中图中变量的指代关系如表 1 所示.

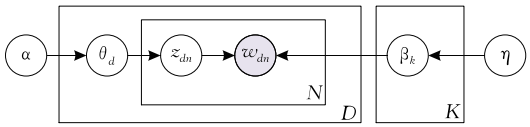

图 2　LDA 图模型

在图 2 中，方框右下角数字是方框内包含变量的重复次数；灰色实心圆表示观测值，空心圆表示隐含随机变量；箭头表示变量的依赖关系；变量之间的指代关系如表 1 所示.

表 1　LDA 变量符号标注

符号	解释
K	表示主题个数
D	表示文档个数
N	表示文档包含的单词个数
w_{dn}	表示第 d 篇文档的第 n 个单词
z_{dn}	表示第 d 篇文档的第 n 个主题
$\boldsymbol{\theta}_d$	服从超参数为 α 的 Dirichlet 概率分布
$\boldsymbol{\beta}_k$	服从超参数为 η 的 Dirichlet 概率分布

在图 2 中，w 是可观测变量，其它均为隐变量. 假设一个语料库是 D 篇文档集合；其中，每篇文档 d 是 N 个单词的序列，表示为 $W=[w_1,w_2,w_3,\cdots,w_n]$. 语料库 D 中的每篇文档的生成过程为

(1) 从主题 Dirichlet 概率分布选取一个主题概率分布 $\boldsymbol{\theta}\sim \mathrm{Dir}(\alpha)$；

(2) 根据得到的主题概率分布中所对应的词概率分布随机采样得到文档中的单词 $p=(\boldsymbol{w}_{d,n}|\boldsymbol{\theta},\boldsymbol{\beta})$；

(3) 重复以上过程，生成文档.

LDA 模型联合概率分布函数如下式所示：

$$p(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{z},w)=\Big(\prod_{i=1}^{k}p(\beta_i|\eta)\Big)\cdot \Big(\prod_{d=1}^{D}p(\boldsymbol{\theta}_d|\alpha)\prod_{n=1}^{N}p(z_{d,n}|\boldsymbol{\theta}_d)\prod_{d=1}^{D}p(z_{d,n}|\beta_{i,k},z_{d,n})\Big) \tag{11}$$

LDA 主题模型由于在词和文档之间加入了主题的概念，使得主题和词汇的概率分布满足 Dirichlet 概率分布，较好地解决了 PLSA 模型在建模过程出现的过拟合问题；但是，在 LDA 模型中假设每个文档的主题概率分布都服从 Dirichlet 概率分布，并没有对不同主题之间的相关性进行刻画，但是，在真实的语料中，不同主题之间存在相关性是一种很普遍的现象；在 LDA 模型中，假设每一篇文档是一个

① The implementation of GibbsLDA is available at https://github.com/asperyang/GibbsLDApy

“词袋”形式，忽略了词序问题；同时，在实践中发现，对于一些短文档，例如微博、评论等，文本数据存在较为严重的稀疏性，直接使用 LDA 建模效果往往不明显；此外，当下流行的在线文本具有明显的时间属性，LDA 模型没有对此进行考虑.

3.2 相关主题模型

在传统的 LDA 模型中，假设主题服从 Dirichlet 分布，并且每个主题之间相互独立. 然而在实际文本语料中，主题之间往往存在一定的相关性. 因此针对 LDA 主题模型这一缺陷，一系列的相关主题模型相继提出，主要包括相关主题模型、四层的 PAM 模型(Pachinko Allocation Model，PAM)、相关高斯主题模型，这三类主题模型都属于无监督的主题模型.

3.2.1 CTM 模型

Blei等人在 2005 年提出了CTM模型(Correlated Topic Model，CTM)[4]，该模型的主要思想是：使用逻辑斯蒂-正态概率分布(Logistic Normal)来代替 LDA 模型中主题的 Dirichlet 概率分布，通过引入协方差矩阵来对主题之间的相关性进行建模. 利用 CTM 模型①生成文档中词汇的图模型如图 3 所示，变量的指代关系如表 2 所示.

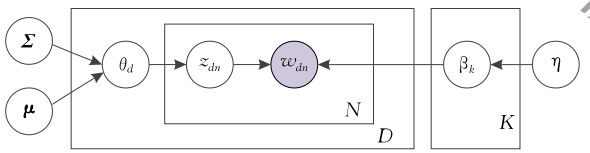

图 3 CTM 图模型

表 2 CTM 变量符号标注

符号	解释
θ_d	表示服从逻辑斯蒂-正态概率分布的主题随机变量
$\boldsymbol{\mu}$	表示辑斯蒂-正态概率分布的 K 维均值向量
$\boldsymbol{\Sigma}$	表示一个 $k\times k$ 维的协方差矩阵

其中，$\theta_d\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 表示服从逻辑斯蒂-正态概率分布的主题随机变量，其中，$\boldsymbol{\mu}$ 表示逻辑斯蒂-正态概率分布的 K 维均值向量，$\boldsymbol{\Sigma}$ 表示一个 $k\times k$ 维的协方差矩阵；$\boldsymbol{\beta}_k$ 表示主题中单词服从参数为 η 的 Dirichlet 概率分布；$w_{d,n}$ 为文档 d 中的第 n 个单词. CTM 模型的联合概率分布函数为

$$\rho(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{z},w)=\Big(\prod_{i=1}^{k}\rho(\boldsymbol{\beta}_i\mid\eta)\Big)\cdot\Big(\prod_{d=1}^{D}\rho(\boldsymbol{\theta}_d\mid\boldsymbol{\mu},\boldsymbol{\Sigma})\prod_{n=1}^{N}\rho(\boldsymbol{z}_{d,n}\mid\boldsymbol{\theta}_d)\prod_{d=1}^{D}\rho(w_n\mid\boldsymbol{z}_{d,n},\boldsymbol{\beta}_{1:k})\Big)\tag{12}$$

在 CTM 模型中，选取可以对主题相关性进行描述的逻辑斯蒂-正态概率分布进行建模，然而，该概率分布与多项概率分布存在非共轭关系，因此将加大后验估计难度.

3.2.2 四层的 PAM 模型

另外的一种相关主题模型是四层的 PAM 模型(Pachinko Allocation Model，PAM)[5]，该模型是 Li 等人在 2006 年提出的，使用 DAG(Directed Acyclic Graph)结构对所有主题相关性进行刻画. PAM 模型利用一个有向无环图的结构实现对所有主题间关系的描述. 其图模型如图 4 所示. 该模型是一个四层的层次结构，其中图模型中变量的指代关系如表 3 所示.

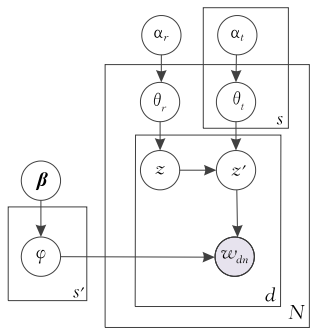

图 4 PAM 图模型

表 3 PAM 变量符号标注

符号	解释
r	表示根节点主题
s	表示第二层级主题，称其为超主题，表示为 $T=\{t_1,t_2,\cdots,t_s\}$
s'	表示第三层级的主题，称为子主题，表示为 $T'=\{t'_1,t'_2,\cdots,t'_{s'}\}$
θ_r	表示根主题(Upper-topic)的多项分布
θ_t	表示子主题(Sub-topics)的多项分布
φ	表示词的多项分布
α_r、α_t 和 $\boldsymbol{\beta}$	分别表示 Dirichlet 概率分布的超参数
w_{dn}	第四层级的叶子节点，表示第 d 篇文档的第 n 个单词

如图 4 所示，根主题节点与所有超主题节点相连，而超主题节点与所有的子主题节点相连，子主题节点与单词节点相连. 其中，假设对于文档 d 的生成过程如下：

(1) $\theta_r^{(d)}$ 是从根节点服从参数为 α_r 的 Dirichlet 概率分布中抽样所得，其中 $\theta_r^{(d)}$ 是关于超主题的一个多项分布；

(2) 对于每一个超主题 t_i，从服从参数为 α_t 的 Dirichlet 概率分布中采样得到 $\theta_{t_i}^{(d)}$，其中 $\theta_{t_i}^{(d)}$ 是关于子主题的一个多项分布；

(3) 对于文档中的每一个单词，首先从多项分布 $\theta_r^{(d)}$ 采样得到一个超主题 z_w，然后从多项分布

① The implementation of CTM is available at https://github.com/kzhai/PyCTM

$\theta_{t_i}^{(d)}$ 采样得到一个子主题 z'_w，进而从分布 $\varphi_{z'_w}$ 中采样得到每一个单词.

因此，一个四层的 PAM 模型生成文档的联合概率分布为

$$\rho(\boldsymbol{\theta},\varphi,\boldsymbol{z},w)=\rho(\boldsymbol{\theta}_r|\alpha_r)\prod_{i=1}^{s}\rho(\boldsymbol{\theta}_t|\alpha_t)\prod_{n=1}^{N}\rho(\boldsymbol{z}_{dn}|\boldsymbol{\theta}_r^{(d)})\rho(\boldsymbol{z}'_{dn}|\boldsymbol{\theta}_{dn}^{(d)})\rho(w_n|\varphi_{z'}) \tag{13}$$

3.2.3　相关高斯主题模型

研究发现，单词嵌入表示[6]可以很好地抓住语言中的语义信息，在单词嵌入向量空间上，语义关系以及词与词的相关性可以同时获得. 因此，Xun 等人在此基础上提出一种相关高斯主题模型(Correlated Gaussian Topic Model，CGTM)[7]. 在单词嵌入模型中，每个文档都是连续的单词嵌入向量序列，而不是离散的单词类型序列. 由于单词嵌入向量表示是基于单词的语义和语法信息并在语义和语法子空间中的位置形成的[8-9]，因此传统的主题模型将不再适用. 因此把每个主题都描述为向量空间中的多维高斯分布，CGTM 模型的图模型如图 5 所示.

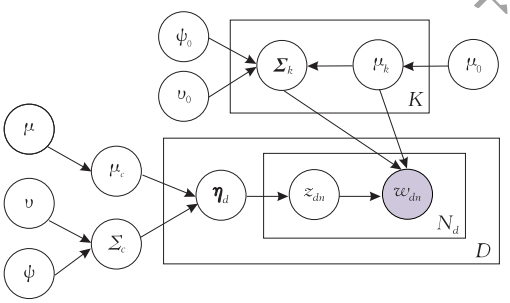

图 5　CGTM 图模型

图 5 表示共有 K 个主题，其中，每个主题表示为单词嵌入向量空间中的多维高斯随机变量，μ_k 和 $\boldsymbol{\Sigma}_k$ 表示第 k 个高斯主题随机变量的均值和协方差；每篇文档由 K 个主题组成.

其中，$\boldsymbol{\eta}_d$ 是一个 k 维向量，每一维表示每个主题在文档 d 中的权重，因此，每篇文档的具体的主题分布是由 $\boldsymbol{\eta}_d$ 所获得；μ_c 和 $\boldsymbol{\Sigma}_c$ 表示 $\boldsymbol{\eta}_d$ 的均值和协方差；通过使用逻辑斯蒂-正态先验来替代传统 LDA 模型中的 Dirichlet 先验，进而能够获取主题之间的相关性. μ、υ、ψ、μ_0、υ_0 以及 ψ_0 为相应的逻辑斯蒂-正态先验和高斯主题的超参数.

该模型的生成过程如下所示：

(1) 获取 $\boldsymbol{\Sigma}_c\sim\mathcal{W}^{-1}(\Psi,\upsilon)$；

(2) 获取 $\mu_c\sim\mathcal{N}\left(\mu,\frac{1}{\tau_c}\boldsymbol{\Sigma}_c\right)$；

(3) 对于每个高斯主题 $k=1,2,\cdots,K$：获取主题的协方差矩阵 $\boldsymbol{\Sigma}_k\sim\mathcal{W}^{-1}(\Psi_0,\upsilon_0)$；获取主题的均值 $\mu_k\sim\mathcal{W}^{-1}\left(\mu_0,\frac{1}{\tau}\boldsymbol{\Sigma}_k\right)$；

(4) 对于每一篇文档 $d=1,2,\cdots,D$：获取 $\mu_d\sim\mathcal{N}(\mu_0,\boldsymbol{\Sigma}_c)$；然后对于每一篇文档的索引 $n=1,2,\cdots,N_d$：获取主题 $z_{dn}\sim\text{Multinomial}(f(\mu_d))$，进而获取单词 $w_{dn}\sim\mathcal{N}(\mu_{zdn},\boldsymbol{\Sigma}_{zdn})$.

其中，τ 和 τ_c 是常数因子，$f(\eta)$ 表示逻辑变换：

$$f(\boldsymbol{\eta}_d^k)=\theta_d^k=\frac{\exp(\boldsymbol{\eta}_d^k)}{\sum_i\exp(\boldsymbol{\eta}_d^i)} \tag{14}$$

使用以下共轭先验作为主题参数：其中，$\mathcal{N}$ 表示高斯分布，$\mathcal{W}^{-1}$ 表示 Wishart 分布的逆，但是，注意逻辑斯蒂-正态分布和多项分布之间仍然存在一个非共轭问题，为此需使用数据扩充技术来缓解该问题.

CGTM 模型的联合概率分布如下所示：

$$\rho(\boldsymbol{\eta}_d,\boldsymbol{z},w)=\prod_{d=1}^{D}\rho(\boldsymbol{\eta}_d|\mu_c,\boldsymbol{\Sigma}_c)\prod_{n=1}^{N}\rho(\boldsymbol{z}_{d,n}|\boldsymbol{\eta}_d)\prod_{d=1}^{D}\rho(w_n|\boldsymbol{z}_{d,n},\mu_k,\boldsymbol{\Sigma}_k) \tag{15}$$

CGTM 模型中将文档中的单词替换为有语义信息的单词嵌入表示形式，将主题建模为单词嵌入向量空间上的多维高斯分布，并在连续高斯主题之间实现对主题相关性的学习. 在该模型中通过使用逻辑斯蒂-正态分布替代传统 LDA 模型中的 Dirichlet 分布来完成对主题和单词之间相关性的描述. 然而，由于逻辑斯蒂-正态分布与多项概率分布之间的非共轭性，常规吉布斯抽样方案无法实现参数的学习，因此，可以采用带有数据增广的吉布斯采样进行模型参数学习.

3.2.4　小结与分析

CTM 模型是在 LDA 模型的基础上发展而来的，它克服了 LDA 模型不能表达主题之间相关性的缺陷. 然而，在 CTM 模型中，只能对成对的主题进行建模，不能实现对全局主题相关性描述；此外，该模型中引入的协方差矩阵参数的个数是主题个数的平方和，随着主题个数的增加，算法复杂度将增大. 针对 CTM 模型只能对两个主题间相关性进行描述的不足，四层的 PAM 模型使用一个 DAG 结构来实现对所有主题相关性进行刻画，然而，PAM 模型的一大缺陷是不能对嵌套的层次结构进行描述. CGTM 模型可以在获取主题之间相关性的基础上，通过单词嵌入向量更好地挖掘单词之间的语义关

系.然而值得注意的是,单词嵌入表示的应用,也给模型引入一定的噪声.此外,传统的 LDA 模型以及 CTM 模型都属于静态主题模型,静态主题模型不能对文本时变性进行显式建模.因此,该类模型对含有时序文本的表现能力受到了限制.

3.3 动态主题模型

在实际应用中,随着数据的爆炸式增长,简单的主题模型已经难以对日益变化的文本进行建模.研究发现,网络媒体的在线文本具有明显的时序性,因此,充分挖掘文本的时态信息,分析文本中主题随着时间演化的规律,可以更为准确地挖掘文本信息.针对静态主题模型不能对文本时变主题进行处理这一缺陷,一系列的动态主题模型相继提出,主要包括动态主题模型(Dynamic Topic Model,DTM)以及在此基础上的扩展模型,具体如下文所示,这几类动态的主题模型都属于无监督的主题模型.

3.3.1 DTM 模型

Blei 等人在 2006 年提出了动态主题模型(Dynamic Topic Model,DTM)[10],该模型的核心是建立了一组概率时间序列模型来分析文档集合中主题的时间演化规律,其中 DTM① 的图模型如图 6 所示.

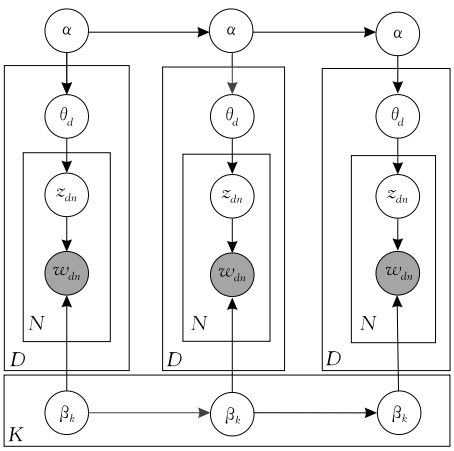

图 6 DTM 图模型

在传统的 LDA 主题分布中,特定文档的主题概率分布 θ 都来自狄利克雷概率分布,在动态主题模型中,使用均值为 α 的逻辑斯蒂-正态分布来表达不确定性程度.模型之间的顺序结构通过一个简单的动态模型来获得:

$$\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I) \tag{16}$$

通过将主题和主题比例分布链接在一起将一组主题模型顺序地绑定在一起.其序列语料库中 t 切片的生成过程如下:

(1) 获取文档的主题 $\beta_t \mid \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \delta^2 I)$;

(2) 获取 $\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$;

(3) 对于每一篇文档:获取 $\eta \sim \mathcal{N}(\alpha_t, a^2 I)$;

(4) 对于每个单词,首先获取主题 $Z \sim$ Multinomial $(\pi(\eta))$,进而获取单词 $W_{tdn} \sim$ Multinomial$(\pi(\beta_{tz}))$.

在此,需注意 π 映射多项分布参数为均值参数,即 $\pi(\beta_{k,t})_w = \dfrac{\exp(\beta_{k,t,w})}{\sum_w \exp(\beta_{k,t,w})}$.

当移除掉水平箭头时,去除了时间动态依赖关系,动态时间模型就变成了一组独立的主题模型.随着时间动态演化的发展,第 t 层的第 k 个主题是从第 $t-1$ 层的第 k 个主题平稳演化而来的.

DTM 的优势在于能够结合文本的时间属性进行建模,极大可能地获取文本随时间演化的规律.然而,DTM 虽然在对含有时序信息的在线文本建模方面取得了成功,但是在实际应用中还面临着许多挑战,例如,DTM 中直接对时间离散处理,忽视了概念漂移产生的偏差;同时,DTM 也存在如何寻找最优时间切片方式的问题.

3.3.2 On-Line LDA

在线文本具有明显的时间属性,针对动态文本流建模问题,Alsumait 等人提出一种在线 LDA 模型(On-Line LDA)[11],当有新的文本流更新的时候,该模型可以利用已得出的主题模型增量式地更新当前模型,不再需要重新访问之前所有的数据,能够实时获取随时间变化的主题结构.

为了阐明这个问题,首先假设文档按照发布日期的升序到达,预定时段大小 ε,例如,一个小时,一天,或者一年,对每个时间切片 t,形成一连串的文件.$\boldsymbol{S}^t = \{d_1, \cdots, d_{M^t}\}$ 是时间切片 t 中包含的文本个数.时间切片 ε 的大小依赖于模型的预测应用场景的性质以及客户对于数据处理结果描述的精细程度的要求.其中 d_1 表示首先达到的文档,d_{M^t} 表示在数据流中最后到达的文档.时间 t 收到的第 n 个文档中的单词表示为 $\boldsymbol{w}_d^t = \{w_{d_1}^t, \cdots, w_{d_N d}^t\}$.假设数据流中有单词在之前的数据流中没有出现,说明数据流 $\boldsymbol{S}^t$ 引入了单词字典表中新的单词,这一假设对简化矩阵 $\boldsymbol{B}$ 的定义和有关计算具有重要意义.变量的指代关系,如表 4 所示.

① The implementation of DTM is available at https://github.com/blei-lab/dtm

表 4　On-line LDA 变量符号标注

符号	解释
δ	表示滑动窗的尺寸
N_d	表示文档中标记词的数量
S^t	表示在时间 t 到达的文档
M^t	在 S^t 中的文档数
$w_{d_i}^t$	在 t 时刻，与文档 d 中第 i 个标记关联的单词
z_i^t	与 $w_{d_i}^t$ 相关的主题
θ_d^t	在 t 时刻，在文档 d 中的特定主题的多项分布
ϕ_k^t	在 t 时刻，在主题 k 中的单词的多项分布
$\boldsymbol{\alpha}_d^t$	在 t 时刻，在文档 d 中先验的 k 维向量
$\boldsymbol{\beta}_k^t$	在 t 时刻，主题 k 的 V^t 维的先验向量
$\boldsymbol{B}_k^t$	主题 k 的 $V^t\times\delta$ 演化矩阵，列为 $\phi_k^i(i\in t-\delta,\cdots,t)$
$\boldsymbol{w}^\delta$	ϕ^i 的权重的 δ 向量$(i\in t-\delta,\cdots,t)$

令 $\boldsymbol{B}_k^{t^{-1}}$ 表示主题 k 的演化矩阵，其中每列 ϕ_k^j 表示在时间 t 特定主题下的单词的概率分布，该矩阵可以看作是在特定时间内文本数据流通过滑动窗口所生成的，例如由 $j\in\{t-\delta-1,\cdots,t-1\}$ 形式. $\boldsymbol{w}^\delta$ 是向量权重，与时间切片的数据流有关，其中假设在 $w^{t^{-1}}$ 时权重加和为 1. 因此，主题 k 在 t 时刻的参数是由主题过去分布的加权组合决定的：

$$\boldsymbol{\beta}_k^t=\boldsymbol{B}_k^{t^{-1}}\boldsymbol{w}^\delta \tag{17}$$

以这种方式计算在连续的模型中 β 的相关的主题分布，进而获取在连续的语料中的主题的演化过程. 因此，在线的 LDA 模型的生成过程如下所示：

(1) 对于每一个主题 $k=1,2,\cdots,K$；

(2) 计算 $\boldsymbol{\beta}_k^t=\boldsymbol{B}_k^{t^{-1}}\boldsymbol{w}^\delta$；

(3) 获取 $\phi_k^t\sim\mathrm{Dir}(\cdot\mid\boldsymbol{\beta}_k^t)$；

(4) 对于文档 d 中的每一个单词标记 w_i：从多项分布 θ_d^t 获取 $z_i(p(z_i\mid\alpha^t))$；从多项分布 ϕ_{z_i} 获取 $w_i(p(w_i\mid z_i,\beta_{z_i}^t))$.

但是该模型使用离散的时间方式，因此灵活性低.

3.3.3　cDTM

DTM 模型的一个主要的弊端是时间的离散化，即如果这个时间粒度选择过大，那么假设文档在一个时间步长内可以交换显然是不正确的；如果时间粒度选择过小，那么随着时间点的增加，变分参数的数量也会激增. 因此，Wang 等人在此基础上，提出了 cDTM(continuous time Dynamic Topic Model)[12]，用于对任意粒度的时序数据进行建模. 该模型的主要思想是通过在语料库中引入布朗运动[13]为主题的演化过程进行显式建模. 其中 cDTM 的概率图模型如图 7 所示.

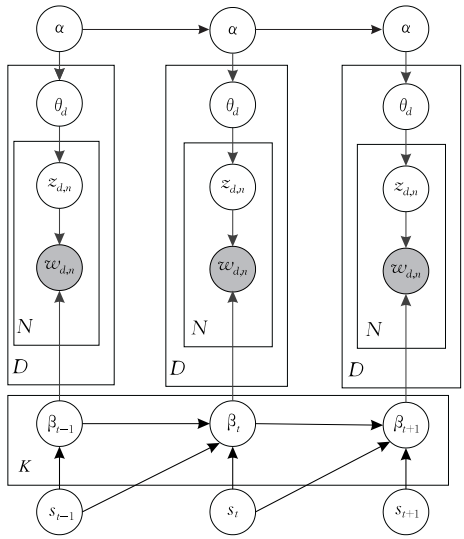

图 7　cDTM 图模型

在 cDTM 中，假设主题参数 β_t 的演变过程服从布朗运动，以此来模拟它们随时间的演化，其中，文档 d_t 的时间戳记为 s_t. $i,j(j>i>0)$ 是两个任意的时间索引，s_i,s_j 是时间戳，$\Delta s_i s_j$ 是它们之间的运行时间. 对于一个具有 K 个主题的 cDTM 模型，第 k 个主题的参数分布表示为

$$\beta_{o,k,w}\sim N(m,v_0) \tag{18}$$

$$\beta_{j,k,w}\mid\beta_{i,k,w,s}\sim N(\beta_{i,k,w},v\Delta s_i s_j) \tag{19}$$

模型的生成过程如下所示：

(1) 对于每个主题 k，$1\le k\le K$：

① 获取 $\beta_{o,k}\sim N(m,v_0 I)$

(2) 在时间戳 s_t，对于每篇文档 d_t：

① 对于每个主题 k，$1\le k\le K$：

从布朗运动模型，获取：

$$\beta_{t,k}\mid\beta_{t-1,k,s}\sim N(\beta_{t-1,k,s},v\Delta s_t I)$$

② 获取 $\theta_t\sim\mathrm{Dir}(\alpha)$

③ 对于每个单词：

i) 获取 $z_{t,n}\sim\mathrm{Multinomial}(\theta_t)$

ii) 获取 $w_{t,n}\sim\mathrm{Multinomial}(\pi(\beta_{t,z_t,n}))$

其中，函数 π 表示多项分布参数的映射. 其形式如下所示：

$$\pi(\beta_{t,k})_w=\frac{\exp(\beta_{t,k,w})}{\sum_w\exp\beta_{t,k,w}} \tag{20}$$

3.3.4　小结与分析

时态主题模型可以很好地对文本的时间属性进行建模，在社交媒体等更新较快的文本数据领域有较好的应用，但时态主题模型的建模过程通常是无监督学习过程，因此模型学习的主题可解释性低，有时往往难以理解. 此外，该类模型还面临的一个重要

问题是文档主题数目的确定，因为该类模型通常在文档生成之前，首先对其主题数目进行预先固定，这就表示模型开始训练之前就已知其主题的个数，这往往不切实际的，因此存在一定的偏差.

3.4 监督主题模型

LDA 本质上是一种无监督的机器学习模型，忽略了跟文本相关的一些类别信息. 因此为了解决机器学习中有监督学习分类的问题，基于监督学习的主题模型开始流行，主要包括监督主题模型，以及在此基础上的扩展模型.

3.4.1 Supervised LDA

Blei 等人在 2007 年提出了监督主题模型(supervised Latent Dirichlet Allocation, sLDA)[14]. 监督主题模型的假设与 LDA 基本相同，不同之处是监督主题模型多了一个关于预测变量的假设，sLDA 模型为每篇文档关联一个服从正态分布的实值响应变量(Response variable)，代表每篇文档的类别标识. sLDA 也可以使用其它类型的响应变量，例如无约束实值、约束为正的实值、有序或无序类标签、非负整数等. sLDA① 生成过程的图模型如图 8 所示.

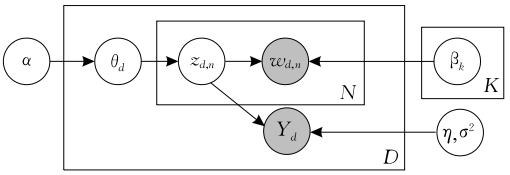

图 8 sLDA 图模型

其中，$\boldsymbol{Y}_d$ 为服从正态分布的响应变量，η、σ^2 为响应正态分布的参数. sLDA 模型的联合概率分布如下所示：

$$\rho(\theta_d, z_d, y, w|\alpha,\beta,\mu,\delta^2)=\prod_{d=1}^{D}\rho(\theta_d|\alpha)\left(\prod_{n=1}^{N}\rho(z_{d,n}|\theta_d)\rho(w_{d,n}|z_{d,n},\beta)\rho(y_d|\boldsymbol{\eta}^{\mathrm{T}})\right) \tag{21}$$

该模型中 α、β、η、σ^2 为未知参数，与无监督的 LDA 模型类似，可以使用变分 EM 算法完成参数学习.

3.4.2 Discriminative LDA

Lacoste-Julien 等人[15]提出一种附加标签的判别主题模型(Discriminative LDA)，在该模型中假设存在监督附加信息(side information)，因此希望考虑该附加信息进而寻找模型的降维表示. 假设每篇文档都与一个分类变量或类标签相关联，即 $y_d=\{1,2,\cdots,C\}$，在主题概率分布上引入类标签独立线性变换. DiscLDA 的概率图模型如图 9 所示.

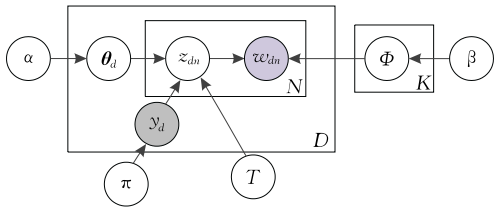

图 9 DiscLDA 图模型

其中，对每一个标签 y，引入一个线性变换 $\boldsymbol{T}^y$：$R^K\rightarrow R^L$，将 K 维 Dirichlet 变量 $\boldsymbol{\theta}_d$ 变换为 Dirichlet 分布的混合向量 $\boldsymbol{T}^y\boldsymbol{\theta}_d\in R^L$，在模型的生成过程中，$\boldsymbol{T}^y\boldsymbol{\theta}_d$ 获取文档的主题概率分布，进而生成 w_n. 通过使用变换矩阵 $\boldsymbol{T}^y$ 对主题进行混合比例的变换作为文档的新表示，最后通过最大化条件似然来估计该模型参数. 然而，DiscLDA 模型与 sLDA 相似，都假设文档只与单一的标签相关联，不能对多标签主题进行分析.

3.4.3 Label-LDA

针对以上文档只与单一的标签相关联的问题，Ramage 等人提出一种标签主题模型[16](Label-LDA)，Label-LDA 模型将文本表示为标签的多项概率分布，有效地解决了文本的多标签判定问题. 其图模型如图 10 所示.

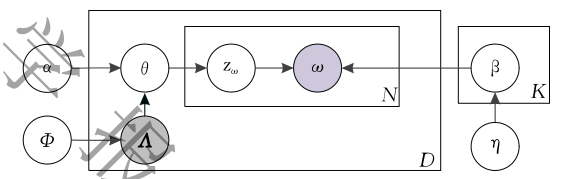

图 10 Label-LDA 图模型

带标签的 LDA 是一个概率图模型，它描述了生成带标签的文档集合的过程. 与隐狄利克雷分配类似，标签 LDA 将每个文档建模为基础主题的混合，并从一个主题生成每个单词. 与 LDA 不同，L-LDA 通过简单地约束主题模型，使之仅使用那些与文档(观察到的)标签集相对应的主题来进行合并监督学习.

其中，假设每一篇文档是由一系列单词 ω 和二值主题指示向量 $\boldsymbol{\Lambda}$ 组成，具体而言，第 d 篇文档的主题指示向量表示为 $\boldsymbol{\Lambda}^{(d)}=(l_1,l_2,\cdots,l_k)$ 一个二值主题指示向量，其中，$l_k\in\{0,1\}$，每个分量代表了主题的存在和不存在两种形式；第 d 篇文档的单词表示为 $\omega^{(d)}=\{w_1,\cdots,w_{N_d}\}$，$w_i\in\{1,\cdots,V\}$ 其中 N_d 表示文档的长度，V 是单词的个数，K 表示语料库中不

① The implementation of SLDA is available at https://github.com/blei-lab/class-slda

同的标签总数. 变量的指代关系，如表 5 所示，模型的生成过程如下所示：

(1) 对于每一个主题 $k\in\{1,\cdots,K\}$：生成 $\boldsymbol{\beta}_k=(\beta_{k,1},\cdots,\beta_{k,V})^{\mathrm{T}}\sim\mathrm{Dir}(\cdot\mid\eta)$；

(2) 对于每一篇文档 d：对于每个主题 k 生成 $\Lambda_k^{(d)}\in\{0,1\}\sim\mathrm{Bernoulli}(\cdot\mid\Phi_k)$；

(3) 生成 $\boldsymbol{\alpha}^{(d)}=\boldsymbol{L}^{(d)}\times\alpha$

(4) 生成 $\boldsymbol{\theta}^{(d)}=(\theta_{l_1},\cdots,\theta_{l_{Md}})^{\mathrm{T}}\sim\mathrm{Dir}(\cdot\mid\boldsymbol{\alpha}^{(d)})$；

对于每一个 i 在 $\{1,\cdots,N_d\}$：生成 $z_i\in\{\lambda_1^{(d)},\cdots,\lambda_{M_d}^{(d)}\}\sim\mathrm{Multinomial}(\cdot\mid\boldsymbol{\theta}^{(d)})$，生成 $w_i\in\{1,\cdots,V\}\sim\mathrm{Multinomial}(\cdot\mid\beta_{z_i})$.

表 5　Label-LDA 变量符号标注

符号	解释
$\boldsymbol{\beta}_k$	表示第 k 个主题的多项分布参数向量
α	表示 Dirichlet 主题先验概率分布参数
η	表示单词的先验概率分布参数
Φ_k	表示第 k 个主题的标签的先验分布参数
$\boldsymbol{\Lambda}$	二值(存在/不存在)主题指示向量
$\boldsymbol{L}^{(d)}$	表示投影矩阵

模型生成过程中，从主题的多项概率分布中生成词的过程与传统的 LDA 是相同的，不同之处是需要生成文档的标签，首先需要生成文档的标签 $\Lambda^{(d)}$，下一步定义文档的标签向量 $\boldsymbol{\lambda}^{(d)}=\{k\mid\Lambda_k^{(d)}=1\}$，在此需要定义一个特定的文档标签投影矩阵 $\boldsymbol{L}^{(d)}$，大小为 $M_d\times K$，其中 $M_d=|\boldsymbol{\lambda}^{(d)}|$，如下式所示，对于每一行 $i\in\{1,\cdots,M_d\}$，每列 $j\in\{1,\cdots,K\}$.

$$L_{ij}^{(d)}=\begin{cases}1, & \lambda_i^{(d)}=j\\ 0, & \text{其它}\end{cases}\tag{22}$$

只有当第 i 个文档的标签 $\lambda_i^{(d)}$ 等价于主题 j，那么 $\boldsymbol{L}^{(d)}$ 的第 i 行，第 j 列为 1，否则为 0. 在此使用 $\boldsymbol{L}^{(d)}$ 矩阵将 Dirichlet 主题先验概率分布参数投影到更低维空间 $\boldsymbol{\alpha}^{(d)}$：

$$\boldsymbol{\alpha}^{(d)}=\boldsymbol{L}^{(d)}\times\alpha=(\alpha_{\lambda_1^{(d)}},\cdots,\alpha_{M_d^{(d)}})^{\mathrm{T}}\tag{23}$$

3.4.4　Multi-annotator sLDA

监督模型的提出，可以有效利用文档中的标签信息. 标签的使用可以更好地对挖掘的主题进行解释，然而使用的标签多数为人为生成，具有一定的主观性. Rodrigues 等人提出一种多注释的监督主题分类模型(Multi-annotator sLDA)[17]，在该模型中分别让多个不同的标注者对文档所属的类别进行标注，进而建模，实现对文档的分类. 该模型可有效地降低人为标注的主观性，降低人为偏差. 其图模型如图 11 所示.

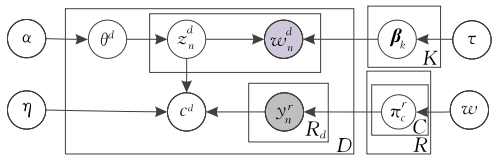

图 11　Multi-annotator sLDA 图模型

其中，$D=\{w^d,\boldsymbol{y}^d\}_{d=1}^{D}$ 表示一个具有注释的文档集，其中对每篇文档中的词 w^d，都有相应的来自 R_d 个不同注释者的标签数据集 $\boldsymbol{y}^d$，其中，$\boldsymbol{y}^d=\{y_r^d\}_{r=1}^{R_d}$. 模型使用噪声模型对来自不同注释者的标签进行建模，给定一个正确的标签类别 c，来自不同注释者 r 所标注的标签表示为 l，产生的概率表示为 $\pi_{c,l}^r$，通过对不同注释者的 $\boldsymbol{\pi}^r$ 矩阵进行建模，以此来降低来自具有不同专业背景标注者的偏差. 变量的指代关系，如表 6 所示，该模型的生成过程如下所示：

(1) 对于每个注释者 r 以及对于每个类别 c：获取它的可靠的参数 $\boldsymbol{\pi}_c^r\mid\omega\sim\mathrm{Dir}(\omega)$；

(2) 对于每个主题 k：获取主题的分布 $\boldsymbol{\beta}_k\mid\tau\sim\mathrm{Dir}(\tau)$；

(3) 对于每一篇文档 d：获取主题分布 $\theta^d\mid\alpha\sim\mathrm{Dir}(\alpha)$；

(4) 对于第 n 个单词：获取主题多项分布 $z^d\mid\theta^d\sim\mathrm{Multinomial}(\theta^d)$，获取每个单词 $w_n^d\mid z_n^d,\beta\sim\mathrm{Multinomial}(\beta_{z_n^d})$；获取隐真实类标签 $c^d\mid z^d,\eta\sim\mathrm{softmax}(\bar{z}^d,\eta)$；

(5) 对于每个注释者 $r\in R_d$：获取注释者的标签 $y^{d,r}\mid c^d,\boldsymbol{\pi}^r\sim\mathrm{Multinomial}(\pi_{c^d}^r)$.

其中 R_d 表示对于第 d 篇文档的多个注释者的标注，$\bar{z}^d=\dfrac{1}{N_d}\sum_{n=1}^{N_d}z_n^d$，其中的归一化指数函数(softmax)表示为

$$p(c^d\mid z^d,\eta)=\frac{\exp(\boldsymbol{\eta}_c^{\mathrm{T}}\bar{z}^d)}{\sum_{l=1}^{c}\exp(\boldsymbol{\eta}_c^{\mathrm{T}}\bar{z}^d)}\tag{24}$$

表 6　Multi-annotator sLDA 变量符号标注

符号	解释
K	表示文档的主题数
C	表示文档的类别数
R	表示注释者的总数
N_d	表示文档 d 中的单词的数量
$\boldsymbol{\pi}_c^r$	表示来自不同标注者的融合矩阵
y_n^r	表示来自不同注释者的标签数据集
c^d	表示文档 d 的真实的类别标签

为降低人为添加标签的主观性，Mei 等人提出一种无监督自动主题标签模型[18]，该模型将添加标

签问题看作两个权衡因素最优组合问题:包括最小化单词分布之间的 Kullback-Leibler 散度和最大化标签与主题模型之间的互信息.因此,该模型实现了使用一个概率方法自动客观地给主题添加标签.然而,该模型仅使用一个文档集合为所有主题生成一个通用的标签候选列表,因此缺乏一定的实用性.Lau 等人在此基础上,提出一种新的给主题模型自动添加主题标签的方法[19],该模型中首先使用英语维基百科生成一个主题标签候选集,然后再对这些候选集进行排序,进而选出最佳的主题标签.

3.4.5 MedLDA

为了有效地利用文档的元数据信息,Wadsworth 等人提出了狄氏-多项分布回归主题模型(Dirichlet-Multinomial Regression,DMR)[20].DMR 主题模型通过将文档观察到的元数据特征信息,例如作者、出版地点、参考文献和日期等作为文档主题概率分布的逻辑斯蒂-正态分布的先验.有效提高了文本挖掘的准确性.然而在 DMR 模型中使用逻辑斯蒂-正态分布矩阵作为转移矩阵,并不是完全的共轭先验模型,因此在后验推断算法中必须使用数值优化算法,随着主题及标签数量的增加,计算复杂度将加大.

为了有效利用文档的主题和词的元数据信息,并解决 DMR 模型存在的非共轭先验问题,Zhao 等人在 2017 年提出 MetaLDA 模型[21].假设给定一个语料库,传统的 LDA 模型对所有文档主题概率分布使用相同的 Dirichlet 先验,对所有主题单词概率分布使用相同的先验.而在 MedLDA 中,每个文档在其主题概率分布上都有一个特定的 Dirichlet 先验,该先验由文档的元信息计算得到.类似地,每个主题都有一个特定的 Dirichlet 先验,它是由单词元数据信息计算出来的,通过利用数据的标签信息,提高建模精度和主题模型质量.其图模型如图 12 所示.

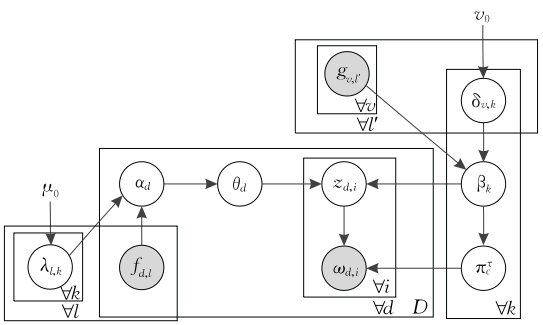

图 12 MedLDA 图模型

在图 12 中,文档的标签是一个二值向量,表示为 $\boldsymbol{f}_d \in \{0,1\}^{L_{doc}}$,其中 L_{doc} 表示不同标签总数;同理,$\boldsymbol{g}_{v,l'}$ 是文档单词的标签信息.$\boldsymbol{\lambda}_{l,k}$ 和 $\boldsymbol{\delta}_{l',k}$ 分别是控制标签 l、l' 对主题和单词影响的参数.其联合概率分布如下所示:

$$p(\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{\phi},\boldsymbol{\beta},w)=p(\boldsymbol{\alpha}_d \mid \boldsymbol{\lambda}_{l,k},\boldsymbol{f}_{d,l})\,p(\boldsymbol{\beta}_k \mid \boldsymbol{\delta}_{l',k},\boldsymbol{g}_{v,l'})\cdot \prod_{d=1}^{D} p(\boldsymbol{\theta}_d \mid \alpha_d) \prod_{k=1}^{K} p(\phi_k \mid \boldsymbol{\beta}_k) \prod_{i=1}^{N} p(\boldsymbol{z}_{d,i} \mid \boldsymbol{\theta}_d) \prod_{d=1}^{D} p(w_{d,i} \mid \boldsymbol{\beta}_{1,k},\boldsymbol{z}_{d,i}) \tag{25}$$

3.4.6 小结与分析

监督主题模型通过增加响应变量,可以有效利用文档的附加信息(side information),例如,评级信息和标签数据等,在处理文本分类问题时优于传统的 LDA 模型;类别 LDA 可以很好识别类间的共享主题以及类间的独立主题;然而这两类模型只能处理单一类别标识文档,不能对多标签主题进行分析,同时,该模型也没有对语料的语义信息进行分析;Lable-LDA 模型实现了类标签与主题之间的一对一关联,然而却不能将标签映射为一个多主题组合,从而容易导致模型与文本之间出现拟合不足的情况.多注释监督模型和自主标签生成模型可有效降低人为生成标签的主观性;MedLDA 主题模型更为有效地利用文档和单词的元数据信息,提高建模的精确性,同时,该模型还具有完美的局部共轭特性,因此可以使用吉布斯采样的方法进行迭代推断,有效减少算法运行时间.

3.5 引入上下文信息

传统的主题模型中假设单词的序列是可以交换的,即“词袋”模型(Bag-of-words),这种模型忽略了文档的结构信息,然而,在实际训练中,常需要考虑文档的上下文信息.因此针对“词袋”模型的缺陷,基于上下文信息的模型相继提出,主要包括 HTMM-LDA、HTMM 模型、MEMMS 模型、CRFs-HTMM 等.

3.5.1 HMM-LDA 模型

Griffiths 等人考虑到隐马尔科夫模型(Hidden Markov Model,HMM)可以获取文档的句法上下文结构信息,而 LDA 模型可以获取文档的语义间关系,因此,提出一种组合模型:HMM-LDA 模型[22].该模型基于由语法和语义约束产生的词之间的不同种类的依赖关系,例如,语法约束导致了相对短时序范围的依赖关系,在一个句子的范围内跨越几个单词,即介词、代词等;语义约束会导致文档的语义相关依赖,即同一文档中的不同句子可能有相似的内

容、使用相似的单词. 在该模型中使用 HMM 模型来对文档中的功能单词(function words)进行处理，使用 LDA 模型来对文档中的语义词汇进行处理，可以提取文档的语法和语义类信息，从而识别单词在文档中扮演的角色，进而实现词性标注或文本分类等任务.

3.5.2　HTMM

在传统的 LDA 模型以及其扩展模型中，都假设主题之间是相互独立的，然而这种强的独立性假设也在一定程度上限制了模型的表达能力. Gruber 等人提出隐主题马尔科夫模型①(Hidden Topic Markov Model，HTMM)[23] 打破了独立性假设，其图模型如图 13 所示，在该模型中，文档中的主题转移满足 Markov 性质，转移概率依赖于 θ，主题的转移变量为 Ψ_n，其中当 $\Psi_n=1$ 时，将表示从 θ 中生成一个新的主题，而当 $\Psi_n=0$ 时，表示第 n 个单词的主题与前一个主题相一致. 在该模型中，假设主题的转移只发生在句子之间，因此，Ψ_n 仅可能是在一个句子的第一个单词处是非零的.

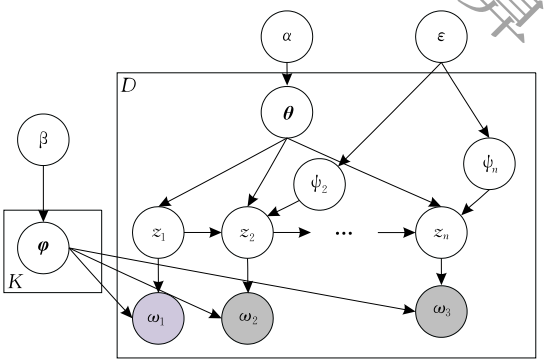

图 13　HTMM 的图模型

HTMM 模型是在 LDA 基础上的改进模型，其概率图模型如上图所示，其中，参数 α，β，$\boldsymbol{\theta}$ 和 $\boldsymbol{\varphi}$ 与传统的 LDA 模型中意义相同；然而在 HTMM 模型中，句子间的主题关系满足 Makov 性质，假设 K 是隐主题数目，N_d 是文档 d 的长度，那么 HTMM 模型生成过程如下所示：

(1) 对于 $z=1,\cdots,K$：获取 $\phi_k\sim\mathrm{Dir}(\beta)$；

(2) 对于 $d=1,\cdots,D$，文档 d 的生成过程如下所示：

① 获取 $\theta\sim\mathrm{Dir}(\alpha)$；

② 设置 $\Psi_1=1$；

③ 对于 $n=2,\cdots,N_d$：

如果是开始的句子，那么 $\Psi_n\sim\mathrm{Binoulli}(\varepsilon)$，其它情况，$\Psi_n=0$；

④ 对于 $n=1,\cdots,N_d$：

i) 如果 $\Psi_n=0$，那么 $z_n=z_{n-1}$，否则 $z_n\sim\mathrm{Multinomial}(\theta)$；

ii) 获取 $w_n\sim\mathrm{Multinomial}(\beta_{z_n})$.

3.5.3　MEMMs

McCallum 等人在 HTMM 模型基础上对其进行改进，在 2000 年提出一种新的马尔科夫序列模型，称其为最大熵的马尔科夫模型(Maximum Entropy Markov Models，MEMMs)[24]. MEMMs 模型允许将观察结果表示为任意的重叠特性(例如单词、大小写、格式、词性)，并定义给定观察序列的状态序列的条件概率. 它是通过使用最大熵框架来拟合一组指数模型来实现这一点的，这些模型代表了给定观测值和先前状态的状态概率. MEMMs 是一个条件概率序列模型，即考虑到相邻状态之间依赖关系，且考虑整个观察序列，因此具有更强的表达能力.

3.5.4　CRFs

Charles 等人在 MEMMs 基础上提出了条件随机场模型(Conditional Random Fields，CRFs)[25] 来构造一个概率模型实现对序列数据进行分割和标注. 其中，MEMMs 和 CRFs 模型最大的区别在于：在 MEMMs 模型中，当给定当前状态以后，使用每一个状态的指数模型作为下一个状态的条件概率；然而，在 CRFs 模型中，给定观测序列以后，有一个单独的指数模型作为整个标签序列的联合概率分布，能够显式地对类标签之间的依赖关系进行建模，因此，在不同状态、不同特征之间的权重可以相互交换.

相较于之前经典的 HMM 模型和基于随即文法的自然语言处理方法，CRFs 模型可以有效放松模型的独立性假设. 此外，基于有向图模型，例如最大熵马尔科夫模型以及其它判别类的马尔科夫模型，对于与父节点状态关联较少的状态容易产生标注偏差，而 CRFs 模型可以有效地避免基于有向图模型的此类限制，缓解标注偏差问题. CRFs 的图模型如图 14 所示.

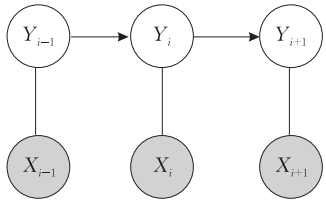

图 14　CRFs 图结构

① The implementation of HTMM and MEMMs is available at https://github.com/NoaKel/NLP-ASS1

在 CRFs 模型中，$\boldsymbol{X}=(\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)$是数据序列上的随机变量，$\boldsymbol{Y}=(\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_v)$是与 $\boldsymbol{X}$ 对应的标记序列上的随机变量. 假设 $\boldsymbol{Y}$ 的所有分量 $\boldsymbol{Y}_i$ 都在一个有限的字符集上取值.

在自然语言场景下，$\boldsymbol{X}$ 表示文本的句子序列，而 $\boldsymbol{Y}$ 表示这些句子的词性标记，而用 $\boldsymbol{Y}'$ 表示这些可能标记的集合. 其中，CRFs 不对随机变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的联合分布建模，而是使用判别框架，通过观测变量和标签序列，构建条件概率模型 $p(\boldsymbol{Y}|\boldsymbol{X})$，没有显式地对边缘概率 $p(\boldsymbol{X})$进行建模.

令 $G=(V,E)$表示一个图，其中 $\boldsymbol{Y}=(\boldsymbol{Y}_v)_{v\in V}$，$\boldsymbol{Y}$ 是图 G 的顶点索引，当以 $\boldsymbol{X}$ 为条件时，$(\boldsymbol{X},\boldsymbol{Y})$是一个条件随机场，随机变量 $\boldsymbol{Y}_v$服从马尔科夫特性：

$$p(\boldsymbol{Y}_v|\boldsymbol{X},\boldsymbol{Y}_w,w\neq v)=p(\boldsymbol{Y}_v|\boldsymbol{X},\boldsymbol{Y}_w,w\sim v),$$

其中 $w\sim v$ 意味着 w 和 v 在图中近邻.

因此，CRF 是基于观测 $\boldsymbol{X}$的全局随机场，在该模型中，默认图 G 是事先确定的. 在对序列进行建模时，$G=(V=\{1,2,\cdots,m\},E=\{(i,i+1)\})$是一个简单的链图，$\boldsymbol{X}$ 可能具有图结构，然而通常没有必要假设 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 具有相同的结构，甚至对于 $\boldsymbol{X}$ 的结构也无需特意来假设.

如果 $\boldsymbol{Y}$ 的图 $G=(V,E)$是一棵树(其中最简单的例子是一条链)，那么边和顶点构成它的图. 因此，根据随机场的基本定理，给定 $\boldsymbol{X}$ 的标记序列 $\boldsymbol{Y}$ 上的联合分布具有这种形式：

$$p_\theta(y|x)\propto \exp\Big(\sum_{e\in E,k}\lambda_k f_k(e,\boldsymbol{y}|_e,x)+\sum_{v\in V,k}\mu_k g_k(v,\boldsymbol{y}|_v,x)\Big) \tag{26}$$

其中 x 是一个数据序列，y 是一个标签序列，$\boldsymbol{y}|_s$是与子图 S 中的顶点相关联的 y 的一组分量.

其中假设给定特征 f_k和 g_k，例如，如果单词 $\boldsymbol{X}_i$是大写字母，标签 $\boldsymbol{Y}_i$是“专有名词”，那么布尔顶点特征 g_k可能为真.

参数的估计问题就是从训练数据 $D=\{(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})\}_{i=1}^N$，使用经验分布 $\tilde{p}(\boldsymbol{x},\boldsymbol{y})$通过最大化以下目标似然函数，确定模型参数 $\boldsymbol{\theta}=(\lambda_1,\lambda_2\cdots;\mu_1,\mu_2\cdots)$：

$$O(\theta)=\sum_{i=1}^N\log p_\theta(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)})\propto\sum_{\boldsymbol{x},\boldsymbol{y}}\tilde{p}(\boldsymbol{x},\boldsymbol{y})\log p_\theta(\boldsymbol{y}|\boldsymbol{x}) \tag{27}$$

3.5.5 小结与分析

在传统模型中，通常假设每一篇文档是一个“词袋”的形式，没有对文档中词序进行考虑. 通过在 LDA 模型中引入隐马尔科夫模型，能有效获取文档的上下文信息；而 HTMM 的文档生成过程中考虑到文档中词的顺序，通过二项概率分布描述句子间主题的转移. 因此，可以更好地挖掘文档的结构信息，HTMM 模型相较于传统 LDA 模型也更具广泛性，同时能更好地实现词义消歧. 然而，HTMM 在文档生成上只表现为相邻句子间的主题转移概率分布，而非真正意义上的全局主题转移概率分布；特征表示有限，不能表示重叠特征(Overlapping)，例如大小写、词缀等等；HTMM 模型假设在同一个句子中的所有单词有共同的主题，假设条件过于严格，可能产生一定的建模偏差；在 MEMM 模型中，虽然加强了模型的表示能力，但是只在局部做归一化，因此容易陷入局部最优；CRFs 模型虽然解决了 MEMM 模型中的标签偏置(label bias)问题，但是这也增加了计算的复杂度，在实际应用中往往由于计算复杂性过高而无法实现.

3.6 贝叶斯非参模型

在传统参数主题模型中，主题个数是事先人为设定的，无法实现自适应学习过程；参数贝叶斯模型在训练过程中，为了降低模型的复杂度，常常假设先验概率分布和后验概率分布是指数族分布，例如，在传统的 LDA 模型中，单词的先验概率分布为 Dirichlet 分布，而后验概率分布是多项分布. 但是这种假设是极具主观性的，在实际中，往往会存在一定的偏差. 因此，为了解决上述问题，贝叶斯非参模型相继提出.

3.6.1 HDP

Teh 等人在 2006 提出基于分层的狄利克雷过程的主题模型[26](Hierarchical Dirichlet Processes，HDP)，该模型的主要组成部分是狄利克雷过程(Dirichlet Processes，DP). HDP 模型①根据文档的集合进行建模，使用预先定义的多层结构，每一层的结果都用一个 DP 表示，每一层中 DP 的概率由上一层决定，自适应地完成对主题个数的学习.

3.6.2 HLDA

传统的 LDA 主题模型是一种线性的主题模型，认为所有主题都在同一层上，即“平的”概率分布，因而无法体现多个主题是如何关联的. 为了利用文本数据的层次结构信息，Blei 等人在 2004 年提出一种分层的主题模型(Hierarchical Topic Model，HLDA)[27]，该模型利用主题树形式来替代原来主题都在同一层的方式，实现对主题的分层处理. 同时，在 HLDA 模型中通过使用非参数贝叶斯模型来

① The implementation of HDP is available at https://github.com/blei-lab/hdp

生成先验概率分布，进而通过模拟“中餐馆”问题来构建主题层次结构. 其图模型如图 15 所示.

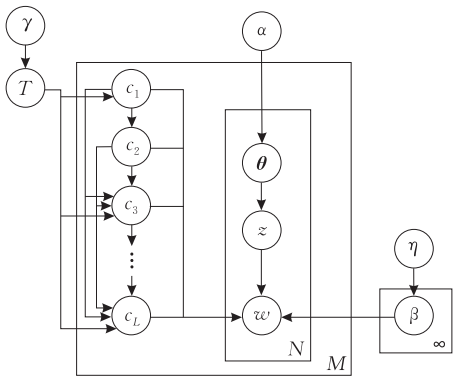

图 15　HLDA 图模型

假设给定一个 L 层树，其中每个节点都跟主题相连，一篇文档的生成过程是：首先从根节点选择一个路径到叶子节点；然后从 L 维的 Dirichlet 分布中，获取主题的比例参数 $\boldsymbol{\theta}$；最后沿着从根节点到子节点的路径，从混合的主题分布 $\boldsymbol{\theta}$ 中，生成文档的单词. 在该模型中，使用嵌套的中国餐馆过程来构建树的层次结构(Chinese Restaurant Process，CRP)，进而放松了事先固定树结构的模型假设.

3.6.3　STM

针对语言的句法问题，Basili 等人[28]提出了一种句法主题模型(Syntactic Topic Model，STM)，也是一种非参数的贝叶斯模型. 该模型的主要优势是在主题选择的过程中，不仅考虑整个文档的主题概率分布，而且还考虑到句法树中父节点的主题类型，分析其主题的转移概率. 因此，在用该模型前，首先对语料进行句法分析得到语法树，进而在模型训练完成后，获取主题个数，同时呈现其语义和句法上的相关性. 其概率图模型如图 16 所示. 在图 16 中，$\boldsymbol{\pi}_k$ 表示在句法树中主题的转移概率；$\boldsymbol{\theta}_d$ 表示每篇文档主题的混合比例；$\boldsymbol{\tau}_k$ 表示在主题 z 下的词概率分布.

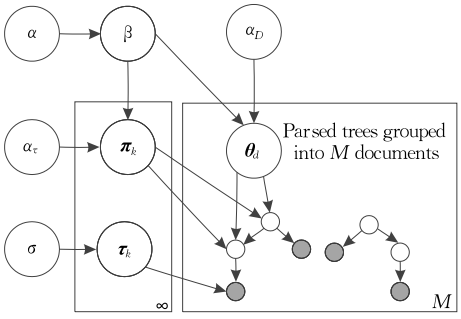

图 16　STM 图模型

Lu 等人考虑到传统关系主题模型假设隐主题个数是作为先验信息人为确定的，这个模型假设通常不符合实际的情况，因此在传统 RTM(Relation Topic Model)上，提出一种非参数的关系主题模型(Nonparametric Relational Topic，NRT)模型[29]，该模型可以自适应对主题个数进行学习.

3.6.4　HPYP

Pitman-Yor 过程是狄利克雷(Dirichlet)过程的一种推广，是一种以概率分布为参数的随机过程. 这些过程可以叠加起来形成一个分层的非参数贝叶斯模型. 因此，Lim 等人提出一种分层 Pitman-Yor 过程(Hierarchical Pitman-Yor Process，HPYP)模型[30]，该模型是经典的 LDA 模型的扩展，通过使用 Pitman-Yor 过程来替代原来的 Dirichlet 分布来实现对文档的建模，变量的指代关系. 如表 7 所示，其图模型如图 17 所示.

表 7　HPYP 主题模型的变量列表

变量	描述
z_{dn}	表示单词 w_{dn} 的主题标签
w_{dn}	表示在第 d 篇文档的第 n 个位置观测到的单词
ϕ_k	表示在第 k 个主题下生成单词的概率分布
θ_d	表示在第 d 篇文档下生成主题的概率分布
γ	表示单词概率分布 ϕ_k 的先验
ν	表示主题概率分布 θ_d 的先验
μ	表示主题 ν 的先验

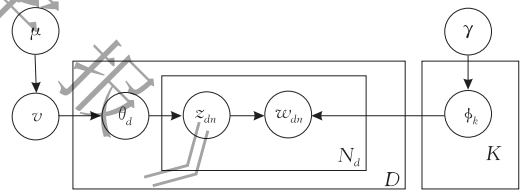

图 17　HPYP 主题模型的图模型

在图 17 中，根节点处 γ 和 μ 两个概率分布是 PYPs 分布：

$$\begin{aligned}\mu &\sim \mathrm{PYP}(\alpha^{\mu},\beta^{\mu},H^{\mu})\\ \gamma &\sim \mathrm{PYP}(\alpha^{\gamma},\beta^{\gamma},H^{\gamma})\end{aligned} \tag{28}$$

其中，在主题模型中，μ 是主题的根节点，γ 是单词的根节点，为了使得模型能够学习任意数量的主题，在此，让基分布(base distribution)μ 和 H^{μ} 表示无限样本的连续分布或离散分布.

在此对于 γ，通常基于文本语料库的词汇量的大小来选择离散的均匀分布，该过程能够给在训练集中没有观察到的单词分配一个很小的概率，平滑了单词概率. 因此，$H^{\gamma}=\left\{\cdots,\frac{1}{|\nu|},\cdots\right\}$，其中 $|\nu|$ 表示文本语料库中包含所有单词的势函数.

对于 HPYP 模型的主题方面，如上已经得到 ν，它表示 μ 的子节点，因此有如下形式：

$$\nu \sim \mathrm{PYP}(\alpha^{\nu}, \beta^{\nu}, \mu) \tag{29}$$

对于大小为 D 的文本语料库中的每个文档 d，文档-主题分布为 θ_d，因此对于每篇文档的主题概率分布表示为

$$\theta_d \sim \mathrm{PYP}(\alpha^{\theta_d}, \beta^{\theta_d}, \nu),\ d=1,\cdots,D \tag{30}$$

对于文档中的每一个单词，文档的每个主题 k 都由模型学习得到，其中主题-单词分布 ϕ_k 中得到与主题相关的单词，ϕ_k 是给定父节点 γ 下的 PYP 概率分布，形式如下：

$$\phi_k \sim \mathrm{PYP}(\alpha^{\phi_k}, \beta^{\phi_k}, \gamma),\ k=1,\cdots,K \tag{31}$$

文档中的每个单词 w_{dn} 是通过 n 来索引(从 1 到 N_d)，隐变量 z_{dn} 和每个单词 w_{dn} 分别由 θ_d 和 ϕ_k 所生成，如下式所示：

$$\begin{aligned} &z_{dn} \mid \theta_d \sim \mathrm{Discrete}(\theta_d), \\ &w_{dn} \mid z_{dn}, \phi \sim \mathrm{Discrete}(\phi_{z_d}),\ n=1,\cdots,N_d \end{aligned} \tag{32}$$

其中，α 和 β 是 PYPs 中的折扣因子和集中度参数，在该 HPYP 模型的生成过程中也常称其为模型的超参数.

3.6.5 小结与分析

HDP 模型中主题个数不用事先定义，而是随着文本的逐渐增加，自适应地增加学习主题的个数，然而相较于传统 LDA 模型，后验推断过程是研究的一大难点所在；HLDA 模型将主题进行层次化处理，使其在语义层面上更适合人类的认知，不用预先定义主题的个数，并且可以从语料中估计出主题个数，并且与 LDA 模型在不中主题数下重复实验得到的最优主题数一致；STM 模型通过构建语法树，实现在语义与句法两方面完成隐变量的学习，对主题进行层次化处理时，由于每一个主题都是关于单词的概率分布，因此 STM 模型对然而较高层级的主题，例如根节点主题，分类不够具体.

3.7 链接主题模型

随着社会媒体的发展使得许多在线网络，例如微博、引用网络等产生了大量的文本内容，因此，发现其潜在的结构并对文本内容进行更深层次的分析是人类研究的一大重要问题. 传统的主题模型利用文档的内容属性对文本进行建模，但是文档内容的不相关性容易产生不准确的主题，影响文本的理解. 融合链接的主题模型可以有效地发现网络文本的潜在的结构，提高主题识别的准确性.

3.7.1 Link-LDA 模型

Cohn 等人在 2000 年提出第一个融合链接和内容的主题模型 Link-PLSA[31]，该模型能够对文档内容和文档集合的相互链接，对网页，研究论文档案等进行联合建模. 由于含有链接和超链接的文档通常由术语和引用组成，因此该模型中，使用 PLSA 模型来对文档的内容术语进行建模；PHITS 模型[32]能够对文献引用的概率因子进行分析，常用于文献的计量分析，因此可用 PHITS 模型来对文档的引用进行建模.

然而，Link-PLSA 模型的参数随文档集的增多呈线性增长，且容易出现过拟合. 因此，在此基础上，Erosheva 等人提出了 Link-LDA 模型[33]，使用 LDA 模型来替代 Link-PLSA 中的 PLSA，避免 PLSA 模型导致的过拟合问题. 其 Link-LDA 模型的图模型如图 18 所示.

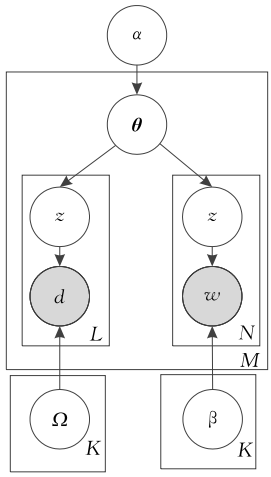

图 18 Link-LDA 模型

在图 18 中，d 表示文档的链接，单词以及链接的生成同 LDA 模型类似，其中两者都具有相同的主题概率分布 $\boldsymbol{\theta}$，来生成各自的潜在主题，因此该模型只能生成文档共享主题下的链接和词.

3.7.2 Pairwise Link-LDA

Link-LDA 和 Link-PLSA 模型定义超链接作为一个随机变量的取值，换句话说，这些模型生成超链接的方式与 LDA 和 LSA 模型生成单词的方式完全相同，因此，这种方式将无法对文档的引用和被引用的主题关系进行显式建模. Nallapati 等人[34]在 2008 年提出 Pairwise Link-LDA 概率模型. Pairwise Link-LDA 模型用 LDA 对每篇文档中单词的生成过程建模，用 MMSB 模型(Mixed Member Ship Stochastic Block，MMSB)[35]来对每对文档中是否生成链接进行建模，联合 LDA 和 MMSB 结构将允许文档对任意的链接结构进行建模，其中，根据时间戳关系完成对每对文档的引用与被引用关系建模. 其图模型如图 19 所示.

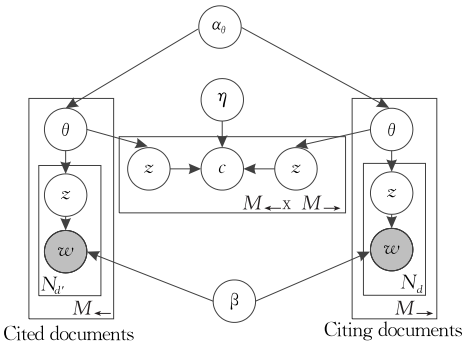

图 19　Pairwise Link-LDA 图模型

在图 19 中，对于一个文档对(d,d')，首先在主题概率分布$\boldsymbol{\theta}_d$中采样得到文档 d 的主题$z_{d,d'}$，类似地，在$\boldsymbol{\theta}_{d'}$中采样得到文档 d'的主题$z_{d',d}$. 文档对之间引用与被引用之间的关系使用一个随机的二进制变量表示，其中$\eta_{z_{d,d'},z_{d',d}}$表示 Bernoulli 概率分布的参数. 与 MMSB 模型不同，Pairwise Link-LDA 模型基于文档的时间戳信息，考虑到链接引用是有向的，因此不具备对称性. 这也将直接导致对大规模文档而言，算法的复杂度将直接加剧；此外，该模型涉及到对每一对文档之间的引用(链接)的存在或不存在进行建模，因此在计算成本上相对较大.

3.7.3　RTM

Zhang 等人提出一种关系主题模型(Relation Topic Model，RTM)[36]，该模型对文档和它们之间的链接进行建模，对于每一个文档对，RTM 模型根据它们的内容将相关的链接建模作为一个二值的随机变量，该模型可以很好地对文档的网络结构进行总结，并且对它们之间的链接关系和文本单词进行预测. 其图模型如图 20 所示.

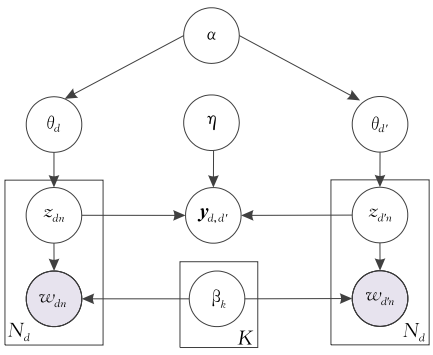

图 20　RTM 图模型

在图 20 中，RTM 模型中文档也是首先从生成主题开始，这个跟传统的 LDA 模型类似. 然而在该模型中，每个文档之间的链接关系通过使用二进制变量$\boldsymbol{y}_{d,d'}$进行建模，最后使用变分推断算法完成对后验参数的估计. 模型的生成过程如下所示：

(1) 对于每篇文档 d：

① 获取主题分布$\theta_d|\alpha\sim\mathrm{Dir}(\alpha)$；

② 对于每个单词$w_{d,n}$：获取特定文档的主题$z_{d,n}|\theta_d\sim\mathrm{Multinomial}(\theta_d)$；获取每个单词$w_{d,n}|z_{d,n}$，$\beta_{1:k}\sim\mathrm{Multinomial}(\beta_{z_{d,n}})$；

(2) 对于每一个文档对 d,d'：

① 获取其二值的链接指示条件概率：$y|z_d,z_{d'}\sim\Psi(\cdot|z_d,z_{d'})$

其中 Ψ 是链接概率函数，定义了两个文档之间的链接分布.

在 2014 年，Taskar 等人提出一种 RTM 模型的扩展模型判别关系主题模型(Discriminative RTM)[37]，该模型通过使用一个规范的贝叶斯推断算法来对主题间相互关系进行建模，呈现了一个基于马尔科夫网络(Markov networks)的可供选择的概率框架. Wang 等人提出一个结合深度学习的关系深度学习模型(Relation Deep Learning，RDL)[38]，该模型对高维的节点属性和具有潜在变量层的链接结构同时建模，针对模型中多个非线性变换问题，文中使用一种广义的变分推断算法来完成对主题的学习以及链接的预测. Terragni 等人考虑到 RTM 已广泛用于发现网络文档集合中的隐藏主题，在此基础上，提出一种基于约束的关系主题模型(Constrained Relational Topic Models，CRTM)[39]. 该模型是 RTM 的半监督扩展，它除了对文档网络的结构建模外，还显式地对一些可用的领域知识建模，即通过使用文档约束的形式结合先验知识，平滑了模型的主题概率分布，进而提高模型的分类准确性.

3.7.4　小结与分析

Link-LDA 模型可以有效避免 Link-PLSA 模型所产生的过拟合问题，在主题挖掘、内容预测以及链接预测方面有较好的应用. 然而，该模型无法对文本链接中的引用与被引用之间的关系建模；Pairwise Link-LDA 模型可以完成对文档引用和被引用关系的建模；然而，由于该模型需要对每对文档之间存在或不存在链接进行显式建模，因此随着文档规模增加，计算量将急剧增加，因此可扩展性低；Link-PLSA-LDA 模型在 Pairwise Link-LDA 模型的基础上有效降低对大规模文档的算法的复杂度；RTM 是一种新的概率生成模型，它可以很好地被用来分析链接集，其扩展模型 Discriminative RTM 的提出有效提高了分类的准确性. 然而，关系主题模型会随着文本数量增加，计算的复杂度会逐渐增加，算法效率降低. 同时，在 RTM 模型和 Dis-RTM 模型中，如何

平衡链接结构和节点属性的学习，是研究的难点所在.

3.8 情感主题模型

情感分析，即识别和提取给定的文本语义的取向，从大量的文本信息中挖掘出有效的情感信息，对信息的提取具有重要的意义.然而，传统的监督模型在情感分析，例如对评论信息的感情色彩分析等方面存在一定的缺陷，因此提出基于情感分析的 LDA 扩展模型具有重要的现实意义.

3.8.1 MG-LDA

Titov 等人在 2008 年提出了多粒度的主题模型(Multi-grain LDA，MG-LDA)[40]，MG-LDA 模型更适用于从在线用户评论中来获取评论对象的可评级方面信息，因为传统的主题模型中获取的是文档的全局隐主题而并非文档中目标对象的用户评级方面信息.MG-LDA 模型不仅可以提取对象的可评级方面信息，而且还可以将它们聚类在一起形成一致的主题.该模型通过对两个不同类型的全局主题和局部主题进行建模，单词可以在全局主题中生成，也可以在局部主题中生成，例如旅游信息评论中，全局主题中包含的单词：三亚酒店，海滨度假；局部主题对应特征，例如酒店位置等.MG-LDA 模型从局部主题中获取用户在线评论的刻面评分等级(aspect rating)，从全局主题中获取具体的属性(property).其概率图模型如图 21 所示.

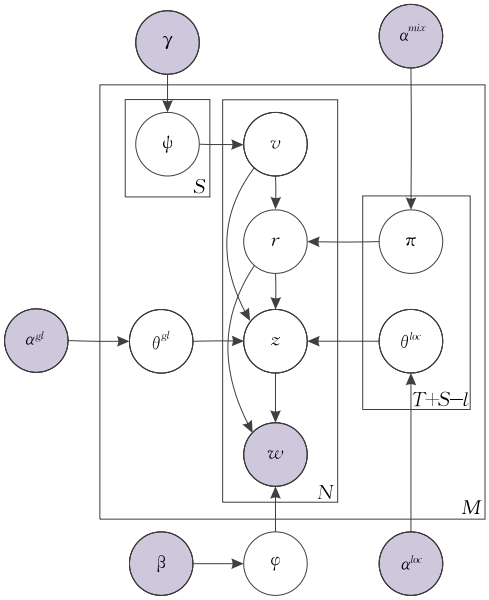

图 21 MG-LDA 图模型

在图 21 中，文档中的句子按一组滑动窗口划分，其中，每个窗口包括文档中的 T 个相邻句子；对于文档 d 中的每个窗口 v，通过其具有的一个与局部主题 $\boldsymbol{\theta}_{d,v}^{loc}$ 相关的概率分布 $\boldsymbol{\pi}_{d,v}$ 来决定窗口中的主题是局部主题还是全局主题，其中，$\boldsymbol{\pi}_{d,v}$ 是一个以 α^{mix} 为超参数的服从 Beta 概率分布的随机变量；

文档的生成过程如下所示：

(1)对于一篇文档 d：首先选取文档的全局主题概率分布 $\boldsymbol{\theta}_{d}^{loc}$，为文档的每个句子选取对应的滑动窗口 $\boldsymbol{\psi}_{d,s}$；对于每个滑动窗口 v，选取对应的局部主题概率分布 $\boldsymbol{\theta}_{d,v}^{loc}$ 以及相关概率分布 $\boldsymbol{\pi}_{d,v}$；

(2)对于句子 s 中的每个词 i，选取对应窗口的下界 $v_{d,i}$ 和上界 $r_{d,i}$；之后生成词对应的全局主题或者局部主题 $z_{d,i}$；

(3)最后，在主题对应的文档的词概率分布 $\varphi_{z_{d,i}}^{r_{d,i}}$ 中生成相应的单词.该模型可以使用吉布斯采样算法完成对参数的估计.

3.8.2 多刻面情感模型(MAS)

Zhu 等人在此基础上，提出了一种 MG-LDA 的扩展模型，称为多刻面情感模型(Multi-Aspect Sentiment，MAS)[41]，该模型从 MG-LDA 模型中获取刻面评分，同时还将每个属性的评分作为观测值加入到该模型中，并假定对属性讨论的文本是对该属性评分的预测信息，将所需要的属性和主题关联起来，因此，该模型是一种监督的主题模型.

3.8.3 JST 模型

Williamson 等人提出主题情感混合模型(Topic Sentiment Mixture，TSM)[42]，该模型把单词分为两大类，一类是与主题无关的功能词汇；另一类是与主题有关的词汇.同时，与主题有关的词汇又分为中性，正面和负面三类，单词的生成过程通过概率在这四大类中选择类，进而在类中选择单词.TSM 模型可以同时获取文档中多个主题以及情感类型.然而，该模型并非直接对情感词汇直接进行建模，而是通过一系列后处理算法来判断情感类型.

TSM 模型是基于 pLSA 模型的扩展模型，为了进一步降低算法复杂度，Lin 等人提出情感-主题联合的模型(Jiont Sentiment Topic model，JST)[43]，该模型是基于 LDA 模型的扩展，通过在文档和主题层之间构建额外的情绪层来实现主题和主题的相关情感信息的联合发现，所以 JST 是四层次的贝叶斯网络模型.其图模型如图 22 所示.

JST 模型①的构造过程如下.假设语料库中有 D 个文档，记为 $\boldsymbol{C}=\{\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_D\}$；每个文档用 N_d 个词的序列表示；去掉语料库中重复的词，剩余的词放入词典中，词典大小为 V，则文档中每个词对应 V

① The implementation of JST available at https://github.com/linron84/JST

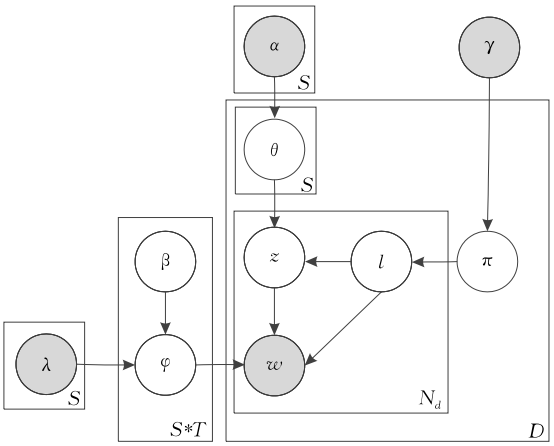

图 22　JST 图模型

中的一个索引项；假设文档主题个数为 T，文档情感个数为 S.

文档中生成词语的过程如下：首先，从带有情感的文档概率分布 $\boldsymbol{\pi}_d$ 中选择一个情感标签 l；之后，在情感标签为 l 的主题概率分布 $\boldsymbol{\theta}_d^l$ 中随机选择一个主题；最后，从带有主题和情感标签对的词语概率分布 $\boldsymbol{\varphi}$ 中生成文档中的词语.

3.8.4　Reverse-JST

Lin 等人在 JST 模型的基础上，提出一种逆 JST 模型（Reverse-JST）[44]，Reverse-JST 同样也是在 LDA 模型的基础上引入一个情感层，构成一个四层的贝叶斯网络，在 JST 模型中，主题的生成依赖于情感标签；而在逆-JST 模型中，情感标签的生成依赖于主题. 在逆-JST 模型中，主题与文档相关，而情感标签与主题相关，单词则与主题和情感标签两者相关. 该模型的生成过程与 JST 模型相似. 其图模型如图 23 所示.

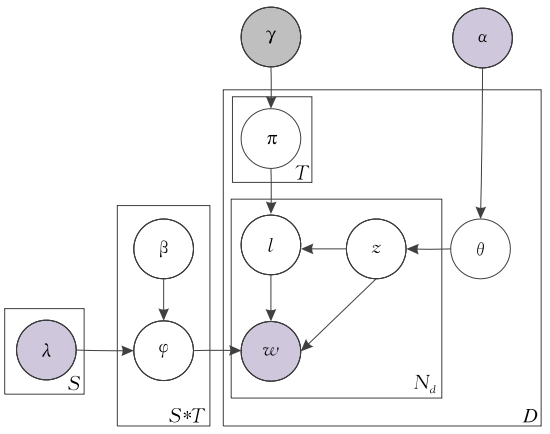

图 23　Reverse-JST 模型

3.8.5　ASUM

Jo 等人针对在线文本评论问题中，如何自动地发现哪些刻面在评论中被评估、以及不同的刻面情感如何表达问题，在 Sentence-LDA 模型的基础上增加了文档的情感概率分布，将刻面和情感结合在一起，形成对不同刻面的情感模型，称其为主题-情感统一的模型（Aspect and Sentiment Unification Model，ASUM）[45]. 其图模型如图 24 所示.

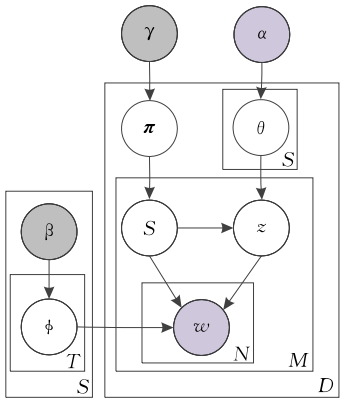

图 24　ASUM 图模型

如图 24 所示，$\boldsymbol{\pi}$ 表示文档 d 中情感概率分布；对于每个情感 s，$\boldsymbol{\theta}_{d,s}$ 表示情感刻面的概率分布；对于文档 d，根据情感刻面对形成文档的单词概率分布为 $\boldsymbol{\phi}_{s,z}$，$\boldsymbol{\phi}_{s,z}$ 的先验信息参数为 β，由于 AUSM 模型中在积极情感词中，消极情感刻面值小；同理，在消极情感词中，积极情感刻面值小，故参数 β 采取非对称取法. 最后，使用吉布斯采样算法完成对 $\boldsymbol{\theta}$，$\boldsymbol{\pi}$ 等隐变量的推断.

3.8.6　SJASM

之前提出的联合情感的主题模型大多是无监督的或者是半监督的模型，因此文本总体的情感分析以及基于整体方面的情感分析在之前的研究工作中，仍未得到解决. 因此，Hai 等人提出一种联合刻面和情感的监督主题模型（a Supervised Joint Aspect and Sentiment Modeling，SJASM）[46]，该模型视文档评论为观点对的形式，可以同时对评论的刻面术语以及相应的观点词汇进行建模. 其图模型如图 25 所示.

在图 25 中，$\boldsymbol{r}_m$ 表示在 d_m 评论中综合评级信息，其中假定 $\boldsymbol{r}_m$ 是在服从参数为 η，δ 的正态线性模型中采样得到；$\boldsymbol{a}$ 和 $\boldsymbol{s}$ 分别为评论 d_m 的刻面和情感的多项概率分布；$\boldsymbol{t}_{mn}$ 和 $\boldsymbol{o}_{mn}$ 分别表示在 d_m 评论中的第 n 个刻面词和观点词，$\boldsymbol{\psi}$ 和 $\boldsymbol{\phi}$ 为相应的刻面词概率分布和意见词概率分布，λ 和 β 为相应的狄利克雷概率分布参数；该模型同时对刻面和相应的观点词同时建模，实现对评论的总体情感分析预测，最后使用吉布斯采样算法实现对 SJASM 模型中参数估计.

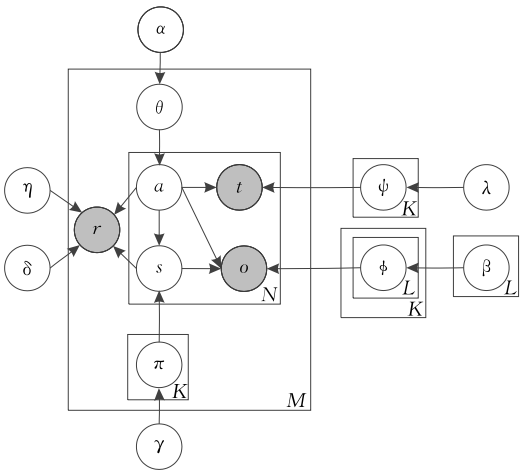

图 25 SJASM 图模型

3.8.7 小结与分析

MG-LDA 模型通过对属性术语出现次数来进行聚类分析,进而获取情感的属性信息.相较于传统的 LDA 模型,MG-LDA 模型是更为合适的对用户的在线评论进行情感分析的方法,然而在该模型中没有对文本的主题与情感之间的联系进行建模;MAS 模型是 MG-LDA 模型的扩展模型,考虑到了主题与情感之间的联系,然而 MAS 模型作为一种监督模型,要求文档中每个刻面必须被评级,这在实际应用中是不可行的;JST 模型以 LDA 为基础,构建了四层的贝叶斯网络,能够无监督地提取文档主题和与之相对应的情感,通过对文本内的词共现信息进行建模,因此在长文本中有较好的应用.然而,JST 模型中假设主题标签和情绪标签在同一层,这与实际情况存在一定的偏差;Reverse-JST 模型与 JST 模型都是通过不同的主题个数来分析不同粒度下的主题和情感的概率分布关系,但是这类模型只考虑了单词的局部情感主题对的概率分布,因此分类效果和模型稳定性容易受局部和主题个数的影响;在线评论学习的句子标签在实际语料中很难训练,然而,ASUM 模型的一个重要的优势是它不需要任何的句子标签就可以建立"句子-主题-词"的三层模型,且细化了情感信息的粒度表达,但该模型认为一个句子中的所有词均属于同一方面,与实际语料不完全相符,假设过于严格;SJASM 模型利用在线评论的总体的情感评分作为监督数据,以此来推断文本的情感刻面以及情感刻面等级,这不但有利于文本的情感分析,而且也可以实现对总体评论的情感预测.

3.9 作者主题模型

在这个信息技术高速发展的时代,随着网络和各种专业数字图书馆的出现,数据量呈爆炸式增长,因此,仅使用单一的主题信息来表示该数据特征已无法满足研究者的需求,更多的是想要挖掘文本主题与其作者间的关联关系,进而能更加全面地发现数据的特征信息.例如,通过对科技文献进行作者主题建模分析,可以有效查找文献领域的相似作者以及构建相关作者研究主题演变图,从而能够更好地挖掘作者与文献间的相关性.基于此考虑,作者的主题模型相继提出.

3.9.1 ATM

Steyvers 等人提出作者主题模型(Author Topic Model,ATM)[47],该模型考虑到作者-文档之间的关系,认为每一个作者应有一个主题概率分布,每一个主题对应有一个主题下单词的概率分布,该模型能够有效地将作者和主题联系起来,在同一框架下,同时在作者级和文档级进行建模.ATM 模型的图模型如图 26 所示.变量的指代关系,如表 8 所示.

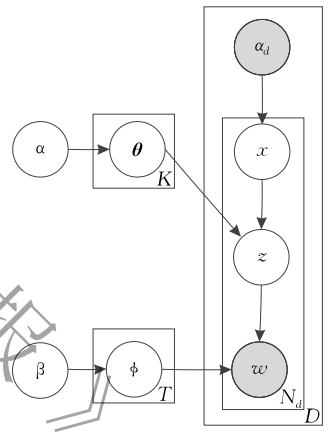

图 26 ATM 图模型

表 8 ATM 变量符号标注

符号	解释
$\boldsymbol{\theta}$	表示作者-主题概率分布
ϕ	表示主题-单词概率分布
$\boldsymbol{a_d}$	表示作者集合上的均匀分布
x	表示文档 d 的作者
z	表示文档 d 的主题
K	表示作者的个数
T	表示主题的个数

在图 26 中,其中,$\boldsymbol{\theta}$ 为作者-主题概率分布;ϕ 为主题-单词概率分布;α 和 β 分别为两者 Dirichlet 先验参数;$\boldsymbol{a_d}$ 为作者集合上的均匀分布;x 为作者;z 为主题;K 为作者的个数;T 是主题的个数.

在文档的生成过程中,先随机选择一个作者,根据作者的主题概率分布生成一个单词,不断地重复该过程,直到生成整个文档.

3.9.2　ACT

自作者主题模型提出以后，基于 ATM 的一系列的扩展模型也被相继提出，针对学术文献，Tang 等人在 2008 年提出作者-会议主题模型（Author Conference Topic Model，ACT）[48]，基于随机游走的框架，对文档、作者及作品的出版场所联合建模，其图模型如图 27 所示. 变量的指代关系如表 9 所示.

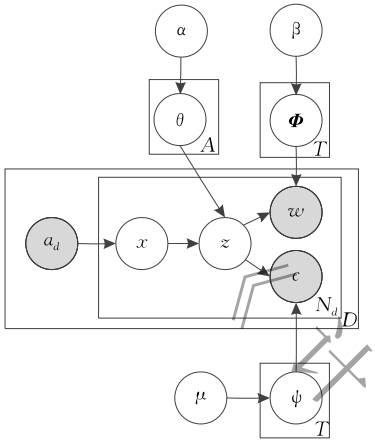

图 27　ACT 图模型

表 9　ACT 变量符号标注

符号	解释
$\boldsymbol{a}_d$	表示文档 d 中作者集合
c	表示文档出版场所
$\boldsymbol{\Phi}$	在主题 z 下所对应的单词概率多项分布
$\boldsymbol{\Psi}$	在主题 z 下所对应的出版场所的多项概率分布

其模型的生成过程如下所示：

（1）对于每一个主题 z，从 Dir(β)和 Dir(μ)中分别获取 ϕ_z 和 Ψ_z；

（2）对于在文档 d 中的每个单词 w_{d_i}：

① 从 $\boldsymbol{a}_d$ 中获取文档的作者 x_{d_i}；

② 对于文档的特定的作者 x_{d_i} 从 Multinomial($\phi_{x_{d_i}}$)中获取文档的主题 z_{d_i}；

③ 从 Multinomial($\Psi_{z_{d_i}}$)中获取单词 w_{d_i}.

3.9.3　ART

针对例如网络电子邮件等有方向性文档问题，McCallum 等人在 2005 年提出作者-接收者主题模型（Author-Recipient-Topic，ART）[49]，与 AT 模型不同的是，ART 模型对于文档的每个主题是由作者和接收者决定的，而不只是单纯的作者，该模型通过发现同一个人作为接受者和发送者两个社会角色的联合主题概率分布，进而确定人物在社会结构中的角色. 其图模型如图 28 所示.

在图 28 中，$\boldsymbol{\alpha}_d$ 表示文档 d 中作者集合，$\boldsymbol{r}_d$ 表示

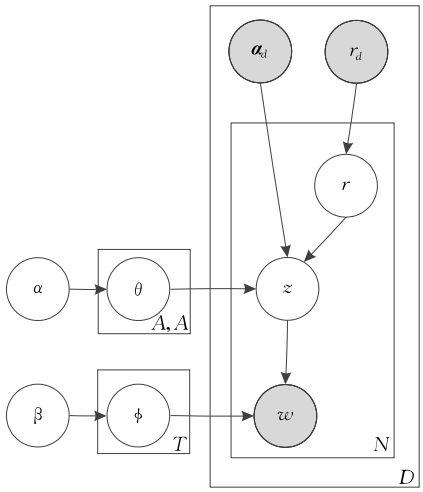

图 28　ART 图模型

相对应的文档接受者. 文档的生成过程如下所示：

为生成文档中的每个单词，通过均匀分布从全部接受者 $\boldsymbol{r}_d$ 中选择对应接受者 r，从特定的作者-接受者对($\boldsymbol{a}_d$，r)决定的概率分布 $\boldsymbol{\theta}_{a_d,r}$ 中获取文档的主题.

3.9.4　TAT

Xu 等人也在 2013 年提出时间作者主题模型（Time Author Topic model，TAT）[50]，在作者主题模型的基础上增加了基于主题的时间概率分布，模型同时结合了文档的作者和时间戳信息. 该模型突破了作者主题模型只能对静态文本建模的局限性，然而该模型还未能扩展到大量的文本集中，因此还不能解决社交媒体等领域中隐社区结构的演化建模问题.

3.9.5　小结与分析

ATM 模型可以通过潜在的主题将文档的作者和词联系起来，有效提高信息检索的准确性，同时，作者-主题结构的出现也为在大量的语料文本中进行重要信息检索提供新的研究思路. 然而，作者-主题模型虽然有一定的创新，但也存在一些问题，例如作者的重名问题、只能应用于静态文本以及无法发现有方向性文档中的主题；ACT 模型通过对作者以及联合出版场所进行建模，可以解决学术文献重名问题，实现名字消歧；ART 模型可以有效挖掘文档背后的人物之间的相互关系，进而确定人物在社会结构中的角色，然而，利用这种邮件的方向性建立人物关系，极易引进噪声，建立不恰当联系，例如垃圾邮件，这也是需要进一步解决的问题.

3.10　词义消歧

词义消歧（Word Sense Disambiguation，WSD）

是自然语言处理研究中一个重要的问题，词义消歧的主要任务是对有歧义的词语，通过结合其背景知识或者上下文信息获取其正确的意思. 词义消歧在机器学习领域，例如机器翻译、信息提取和检索以及问题解答等都有重要的应用. 通常，词义消歧系统是将目标单词周围的句子或者小窗口用作消歧的上下文信息，然而该类方法的复杂度将随上下文的数量增多呈指数级增加. 研究发现，基于主题模型的WSD系统可以将整个文档作为目标单词的上下文来实现词义消歧，有效提高单词语义一致性.

3.10.1 LDAWN

Boyd-Graber 等人首先尝试将 LDA 系统应用于词义消歧中，提出了基于 Wordnet 的 LDA 模型(Latent Dirichlet Allocation with WORDNET, LDAWN)[51]. 在 LDAWN 模型中，作者针对每一个主题定义了一个同义词集(Synset)的转移概率矩阵，该矩阵在生成的过程中，与 LDA 模型类似，根据主题概率分布获取相应主题，不同的是，LDAWN 模型选择一条以实体(Entity)为根节点，不断游历(Walk)直到碰到由单词构成的叶节点，然后输出该单词. 因此，即便单词相同，但由于主题不同，LDAWN 模型在生成该单词时在 WordNet 中选择的路径不同，以此实现词义消歧.

3.10.2 基于知识型的 WSD

Chaplot 等人在 2018 年提出基于知识的单词意思消歧模型(Knowledge-based Word sense Disambiguation)[52]，该模型是 LDA 模型的一种变体，将整个文档作为单词的上下文，使用一个同义词概率分布来替代原来的主题概率分布. Chaplot 等人依据 WordNet 给关于词的同义词概率分布分配一个非均匀的先验信息进行建模，并利用逻辑正态分布，对不同的同义词集的相关性进行描述[52]. 基于知识型的 WSD 模型类似于相关的主题模型，不同之处在于该模型先验知识是固定的而非学习来的. 其图模型如图 29 所示.

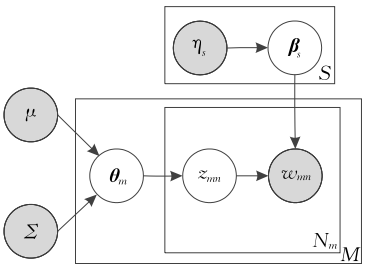

图 29 WSD-LDA 图模型

在图 29 中，$\boldsymbol{s}=\{1,2,\cdots,S\}$ 表示同义词的集合；$\boldsymbol{\beta}_s$ 表示在每一个同义词集合下的词概率分布；$\boldsymbol{\theta}_m$ 是以 μ 和 Σ 为先验信息参数的逻辑正态分布，实现对同义词集的相关性建模. Terragni 等人[39]通过考虑词汇知识库中的图形语义关系来对单词向量表示进行编码，然后对编码的单词向量表示进行相似性度量，进而利用在文本中提取的上下文单词信息实现对歧义词的分析.

3.10.3 小结与分析

在 LDAWN 中，WordNet 层中共享相同路径的词义将被认为具有相同的主题，然而，这样观察到的 WordNet 可能并非 WSD 的最优结构；传统的 WSD 系统[53]通常使用目标词周围的一个句子或者是小窗口来作为单词的上下文实现消歧，然而，该过程会随着上下文的大小呈指数增长. 在 Knowledge-based WSD 系统中计算复杂度随着上下文的增长呈线性关系.

3.11 词向量概率主题模型

基于词向量的概率主题模型主要是通过训练好的词向量来提高主题模型的泛化性能. 短文本往往具有较高的稀疏性，当该类模型应用于短文本学习时，学习的主题词常具有更好的语义一致性.

在没有引入词向量之前，Zhao 等人提出基于狄利克雷多项式的混合模型(Dirichlet Multinomial Mixture, DMM)来推断短文本中隐含的主题[54]. DMM 模型①遵循一个简单的假设，即每个文本只从一个潜在主题中取样而得. 与 LDA 中采用的每个文本都在一组主题上建模的复杂假设相比，这种假设是合理的，并且更适合于短文本. 然而，该类模型忽略的一个重要问题是：不同文本间单词的统计信息往往不具很好的语义一致性，例如，有些单词具有很高相关性，但共现度低.

考虑到短文本通常都缺少足够的词的共现信息，一些模型尝试使用丰富的全局词共现模式来推断潜在的主题，例如直接对短文本中的双词进行建模的 BTM(Biterm Topic Modeling)模型[55]②，在该模型中，假定 Biterm 中的两个词共享同一主题，该主题是从整个语料库的混合主题中抽取的. 由于全局词的共现性，该类模型能在一定程度上缓解了短文本的稀疏性. 此后，基于 BTM 的一系列扩展模

① The implementation of DMM is available at https://github.com/atefm/pDMM

② The implementation of BTM is available at https://github.com/xiaohuiyan/BTM

型[56]被相继提出，例如将 Biterm 的突发性作为先验知识或区分背景词和主题词的主题模型 WNTM (Word Network Topic Model)[57].

此外，针对短文本简短的问题，学者们还提出基于伪文档的主题建模方法，换句话说，就是将多个短文本聚合为一个长文本(伪文档)，然后在该伪文档上进行主题建模. 例如 Quan 等人提出的自聚合主题模型(Self-Aggregation based Topic Model, SATM)[58]①、Shi 等人提出的扩展自聚合动态主题模型[59]、Zuo 等人提出的基于伪文档建模的主题模型 PTM(Pseudo-document-based Topic Model)[60]，均假设每个短文本均采样自某个长文本，且伪文档的所有短文本均包含同一个主题.

自从基于神经网络的词向量分布表示被提出，词向量技术由于其较好的表示学习性能受到研究者的广泛关注[61]. 该方法旨在将词汇表示为低维空间的稠密实值向量，进而能够更好地对词汇间的语义关系进行度量. 在基于词向量的概率主题模型中，各类模型通过使用预训练的词向量来度量词汇之间的语义相似性，进而使得相似的词汇可以对同一主题进行增强，在该过程中，根据其词向量的使用方式的不同分为以下三类：(1) 基于高斯分布的词向量主题模型；(2) 基于词向量增强的主题模型；(3) 基于知识向量的主题模型.

3.11.1　基于高斯分布的词向量主题模型

传统的基于 Dirichlet 先验的主题模型是在“词袋”文档中采样，即从固定规模的词表离散空间中采样. 近年来，随着词向量模型的不断发展，研究者们开始尝试直接在词向量空间中推断主题，此时，传统模型中的 Dirichlet-多项分布假设将不再适用；而对于多元高斯分布，常利用欧式距离来对其词汇间相似程度进行描述，使相似度高的词汇尽量汇聚到同一主题下，提高主题词间的语义一致性.

Das 等人最早尝试从词向量空间中采样主题，提出 GLDA(Gaussian LDA)模型[62]. 在该模型中，假设文档不是由单词类型序列组成的，而是由单词嵌入向量组成的. 由于观测变量不再是离散值，而是多维空间中的连续向量，因此将主题 k 描述为一个均值为 μ_k 和协方差为 $\boldsymbol{\Sigma}_k$ 的多元高斯分布. 其生成过程如下所示：

(1) 对于主题 $k=1$ 到 K：

生成主题多元高斯分布的协方差 $\boldsymbol{\Sigma}_k \sim \mathcal{W}^{-1}(\Psi, \nu)$ 和均值 $\mu_k \sim \mathcal{N}\left(\mu, \frac{1}{\kappa}\boldsymbol{\Sigma}_k\right)$；

(2) 对于语料 D 中的每篇文档 d：

获取其主题分布 $\theta_d \sim \mathrm{Dir}(\alpha)$；

对于文档 d 的第 n 个单词：生成主题分配 $z_n \sim \mathrm{Categorical}(\theta_d)$；根据主题分配 z_n 生成词向量 $\boldsymbol{v}_{d,n} \sim \mathcal{N}(\mu_{z_n}, \Sigma_{z_n})$.

GLDA 模型是基于维基百科语料预训练的词向量，因而，利用该嵌入空间中语义相似的词汇的连续性，即使一个单词以前从未出现过，但它与现存的主题词有相似之处，仍可以分配给该未见单词一个较高的主题概率，提高模型的鲁棒性. 相较与传统的 LDA 模型，该模型也有效地提高了主题词语义一致性.

Li 等人考虑到 GLDA 模型使用欧氏距离并非最优的词嵌入空间的语义度量，因此提出一种新的 MvTM(Mix-von Mises Fisher Topic Model)[63]. 在该模型中，使用混合的 von Mises 分布[64]替换 GLDA中的高斯主题分布，使用词向量的余弦相似度来对词汇间的语义相似性进行度量. 这样，在文档生成过程中，词向量的采样过程满足 $v_{dn} \sim v\mathcal{M}(\Delta_{z_{dn}})$，其中 $\Delta_k=\{\mu_k, \kappa_k\}$ 表示 vMF 分布参数. 相较于 GLDA，该模型提高了模型的分类准确性.

计算成本高及可扩展性差是制约相关主题模型进一步发展的关键因素. He 等人在此基础上提出了一种新的利用分布式表示学习的高斯相关主题模型[65]，该模型通过主题向量之间的紧密性来学习紧凑的主题嵌入并捕获主题间的相关性. 模型的生成过程如下所示.

(1) 对于每个主题 $k=1,2,\cdots,K$：

获取主题单词分布 $\phi_k \sim \mathrm{Dir}(\beta)$；

获取主题向量 $\boldsymbol{u}_k \sim \mathcal{N}(\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})$；

(2) 对于每篇文档 $d=1,2,\cdots,D$：

获取文档向量 $\boldsymbol{a}_d \sim \mathcal{N}(\boldsymbol{0}, \rho^{-1}\boldsymbol{I})$；

获取文档主题权重 $\eta_d \sim \mathcal{N}(\boldsymbol{U}\boldsymbol{a}_d, \tau^{-1}\boldsymbol{I})$；

获得文档-主题分布 $\theta_d=\mathrm{softmax}(\eta_d)$；

(3) 对于每个单词 $n=1,2,\cdots,N_d$

分配主题 $z_{dn} \sim \mathrm{Mutinomal}(\theta_d)$

生成单词 $w_{dn} \sim \mathrm{Mutinomal}(\phi_{z_{dn}})$.

在该模型中引入主题向量 $\boldsymbol{u}_k$ 和文档向量 $\boldsymbol{a}_d$，其中，$\boldsymbol{u}_k$ 和 $\boldsymbol{a}_d$ 分别为服从参数为 α 和 ρ 的高斯分布. 与相关主题模型不同的是，该模型中，文档-主题分布不再从高斯分布中采样，而是根据 $\boldsymbol{u}_k$ 和 $\boldsymbol{a}_d$ 之间的向

① The implementation of SATM is available at https://github.com/WHUIR/SATM

量相似性进行抽样.

3.11.2 基于词向量增强主题模型

基于词向量增强的主题模型在传统模型的基础上,在文本词汇生成的过程中根据事先训练的词向量模型依较大概率将语义相近的词汇分配至同一主题下,进而提高模型的性能.

Nguyen 等人在 2015 年利用一个潜在的特征模型分别与 LDA 和 DMM 相结合[66],提出两个新颖的 LF-LDA(Latent Feature-LDA)模型和 LF-DMM 模型.在 LF-LDA 模型中,引入词向量 $\boldsymbol{v}_w$ 和主题向量 $\boldsymbol{u}_k$,通过计算两者的相似性 $\mu_w=\boldsymbol{u}_k\boldsymbol{v}_w^{\mathrm{T}}$ 作为主题 k 下的单词的权重.在此,定义潜在特征模型为 $CatE(w\mid\mu)\propto\mathrm{softmax}(\mu_w)$. LF-LDA 图模型如图 30 所示.

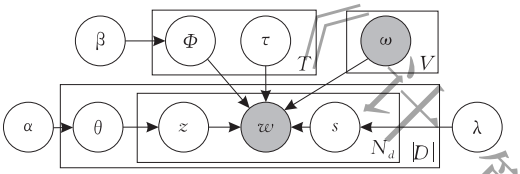

图 30 LF-LDA 图模型

LF-LDA 模型在文档生成过程中,首先,与传统 LDA 模型类似,获取文档 d 的主题序列 $z_d\sim\mathrm{Multinomal}(\theta_d)$;其次,对于第 i 个单词 w_{d_i},从一个伯努利分布中选择一个二进制指示变量 s_{d_i},以此来决定单词 w_{d_i} 是由狄利克雷多项式分布还是潜在特征构件所生成.

其中单词 w_{d_i} 的生成概率表示为 $w_{d_i}\sim(1-s_{d_i})\mathrm{Multinomal}(\phi_{z_{dn}})+s_{d_i}CatE(\boldsymbol{u}_{zd}\boldsymbol{v}_w^{\mathrm{T}})$. LF-DMM 模型的生成过程与此项类似.

Liu 等人针对短文本常面临严重的稀疏性这个问题,在 BTM 模型的基础上,提出一种结合词向量特征的双词主题模型 LF-BTM,与 LF-LDA 模型的建模思想类似,只是基准模型用 BTM 来替换传统的 LDA[67].该模型引入潜在特征模型以利用其丰富的词向量信息来补足短文本内容稀疏问题.

Li 等人在 DMM 模型的基础上,结合 GPU (Generalized Pólya Urn)理论,进而提出 GPU-DMM 模型[68].值得注意的是,该模型仅考虑利用词向量间的语义相似程度来增强其主题的分配概率,而没有在模型中引入其它的例如主题向量等变量,因此,模型具有简洁、高效和易推广等优势.之后,Li 等人在此模型基础上,进一步提出了 GPU-PDMM 模型(Poisson-based GPU-DMM)[69],其中,该模型利用泊松分布来对每篇文档的主题数进行约束,即每篇文档仅包含 1~3 个主题,进而对文档主题数进行了修正,在一定程度上提高模型的分类性能和主题语义连贯性,但同时也增加了时间消耗.

3.11.3 基于知识向量主题模型

传统的概率主题模型,例如 LDA 等,被广泛地应用于文本的建模和分析,然而,这些没有任何人类知识作为先验的无监督模型常常导致难以解释的问题.为了克服主题模型中可解释性的缺陷,研究者们提出将不同形式的领域知识合并到主题模型中,进一步来提升模型的学习能力.

Yao 等人提出了一个新的基于知识图嵌入的主题模型,称为知识图嵌入 LDA(Knowledge Graph Embedding LDA,KGE-LDA)[70].在该模型中,将知识图谱中的实体及向量表示引入 LDA 中,能够显式地对文档级的共现词进行建模,进而能够获取更一致的主题和更好的文档表示.每篇文档 d 包含 N_{w_d} 个单词和 N_{e_d} 个实体,根据 Bordes 等人在 2013 年提出的方法来进行向量表示学习[71].最后,该模型为每个单词和实体各采样一个主题用于文档生成,其中由于实体向量是单位球面上的连续向量,因此主题-实体满足 vMF 分布[64].KGE-LDA 模型利用实体向量对知识进行编码,极大地提高了语义一致性,并能在主题空间中更好地表示文档.

之后,Li 等人在 2019 年提出一种基于知识图嵌入的主题模型(Topic Modeling with Knowledge Graph Embedding,TMKGE)[72],该模型是在贝叶斯非参 HDP 模型的基础上将知识图嵌入到主题建模的上下文中,进而来提取更连贯的主题.

3.11.4 小结

基于词向量的概率主题模型均直接利用事先训练的词向量来辅助模型的学习,使得语义相近的词汇依较大概率获得同一主题,提高主题词的一致性和可解释性,丰富了文本的潜在特征表达,进而有效地提高模型分类的准确性.但是,上述三类模型在前提假设以及词嵌入向量的使用方面存在一定差异.基于高斯分布的词向量主题模型改变了基准模型的假设,采用高斯先验来替换传统模型中的狄利克雷先验假设,然而在进行后验推断时,高斯先验与多项分布不是共轭分布,将直接加大模型求解的复杂度;基于词向量增强的主题模型直接利用词向量与文档向量之间的相似性程度进行建模,进而来提高分类准确性;目前,基于知识向量的主题模型的研究尚处

于起步探索阶段，此外，就其实验效果而言，该模型的文本建模效果一般. 这可能是由于文本稀疏性较高或知识向量抽取的准确性低. 因此，如何进一步提高知识向量抽取的准确性也是未来研究的重要方向之一.

3.12　各主题模型的对比

表 10 对上述的各类主题模型在其模型特点、应用领域、参数学习等方面进行了总结.

表 10　主题模型总结

	模型	监督性	词序性	应用领域	参数学习	说明
	LDA	无	无	情感分析、文献挖掘、社交媒体主题挖掘、句子分割	变分 EM[2]	具有开创性的意义，第一个完整意义上的概率主题模型
相关性	CTM	无	无	新闻、社交媒体中相关性建模	变分 EM[4]	对主题之间的相关性进行描述
	PAM	无	无	文本分类	吉布斯采样[5]	有向无环图的结构实现对所有主题间关系的描述.
	CGTM	无	无	文本分类、对文本相关性建模	吉布斯采样[7]	利用词嵌入对语义关系进行描述
时态性	DTM	无	无	时态文本、历史文献	变分 KL[10]	随时间演化的主题模型
	On-Line LDA	无	无	在线文本、社交媒体	吉布斯采样[11]	当有新的文本流更新的时候，该模型可以利用已得出的主题模型，增量式的更新当前模型
	cDTM	无	无	时态文本	变分 KL[12]	引入布朗运动，连续时间动态主题模型
监督性	sLDA	有	无	文本分类、情感分析	变分推断[14]	单一类别标识文档，进行监督学习
	DiscLDA	有	无	文本分类、文本降维处理	吉布斯采样[15]	附加标签的类别
	Label-LDA	有	无	标签文档聚类	吉布斯采样[16]	每篇文档有若干个标签，解决了文本的多标签判定问题
	Multi-sLDA	有	无	文本分类	变分推断[17]	对不同的标签注释进行建模，降低人为的主观性
	DMR	有	无	摘要提取、有特征文档聚类	吉布斯采样[20]	将文档的元数据信息作为主题概率分布的先验信息
结构性	HMM-LDA	无	有	词性标注、文档分类	MCMC[21,22]	HTMM 获取句法结构和 LDA 来语义关系
	HTMM	无	有	知识发现	EM、向前向后算法[22,23]	以句子为单位分配主题
	MEMMS	无	有	文档结构分析、知识发现	吉布斯采样[23,24]	即考虑相邻状态之间依赖关系，且也对整个观察序列进行考虑
	CRFs	无	有	语义分割，词性标注	EM 算法[24,25]	做归一化时，考虑了全局概率依赖关系
非参性	HDP	无	无	主题挖掘	吉布斯采样[26]	主题个数不用事先定义，自适应学习主题的个数
	HLDA	无	无	主题层次学习问题	吉布斯采样[27]	实现对主题的分层处理
	STM	无	有	文档语法分析、词义消歧	变分推断[28]	不仅考虑整个文档的主题概率分布，而且还考虑到句法树中父节点的主题类型
	HPYP	无	无	主题层次学习问题	吉布斯采样[30]	通过使用 Pitman-Yor 过程来替代原来的经典的 LDA 模型中的 Dirichlet 分布来实现对文档建模.
链接主模型	Link-LDA	有	无	链接发现、学术文献	EM 算法[31]	融合链接和内容的主题模型
	P-Link-LDA	有	无	社区结构发现	变分推断[35]	可以完成对文档引用和被引用关系的建模
	Link-PLSA-LDA	有	无	文献挖掘、网络结构数据挖掘	VB 算法[39]	把文档分成两部分. 用 Link-LDA 生成引用文档内容和链接，用对称 PLSA 生成被引文档的内容
	RTM	有	无	网页引用、网页链接、社区发现	EM 算法[36]	综合考虑了节点属性和它们之间的链接结构
情感主题模型	MG-LDA	无	无	在线用户评论	EM 算法[40]	获取情感的特征信息
	MAS	有	无	在线用户评论，评价情感分析	收敛吉布斯采样[41]	该模型为从 MG-LDA 模型中获取的刻面评分，同时还将每个属性的评分作为观测值加入到该模型中，将所需的属性和主题关联起来.
	JST	无	无	文本情感分析、语义分类	吉布斯采样[42]	能够无监督的提取文档主题和与之相对应的情感
	Reverse-JST	无	无	文本情感分析、语义分类	吉布斯采样[44]	该模型在 LDA 模型的基础上引入一个情感层，构成一个四层的贝叶斯网络.
	ASUM	无	有	在线评论、语义分类	吉布斯采样[45]	以文本句子作为情感分析的最小单位，进一步细化了情感信息的表达粒度
	SJASM	有	有	文本、评论情感分析	吉布斯采样[46]	该模型利用在线评论的总体的情感评分作为监督数据，以此来推断文本的情感方面以及情感方面等级

（续 表）

	模型	监督性	词序性	应用领域	参数学习	说明
作者主题模型	ATM	无	无	文本检索、文献作者建模	吉布斯采样[47]	每一个作者有一个概率主题概率分布
	ACT	无	无	文本检索、文献作者建模	吉布斯采样[48]	基于随机游走的框架，通过对文档、作者及作品的出版场所联合建模.
	ART	无	无	文献检索	吉布斯采样[49]	文档的每个主题是由作者和接收者决定的，通过发现同一个人作为接受者和发送者两个社会角色的联合主题概率分布.
	TAT	无	无	文献检索	吉布斯采样[147]	同时结合文档的作者和时间戳信息
词义消歧	LDAWN	无	无	对文本词义消歧	吉布斯采样[51]	有向无环图的结构实现对所有主题间关系的描述.
	WSD	无	无	对文本词义消歧	吉布斯采样[52]	将整个文档作为单词的上下文，使用一个同义词概率分布来替代原来的主题概率分布
词向量	GLDA	无	无	文本分类、主题发现	坍缩吉布斯采样[62]	文档不是由单词类型序列组成的，而是由单词向量拼接而成，直接词向量空间中采样主题
	MvTM	无	无	文本分类、主题发现	坍缩吉布斯采样[63]	使用混合的 vMF 分布替换 GLDA 中的高斯主题分布，使用词向量的余弦相似度来对词汇间的语义相似性进行度量
	GCTM	无	无	文本分类	随机变分[65]	在词向量空间中采样，能够捕获主题间的相关性
	LF-LDA	无	无	文档分类、文档聚类	吉布斯采样[66]	在词向量空间中采样，能够捕获主题间的相关性
	LF-DMM	无	无	文档分类、文档聚类	吉布斯采样[66]	根据主题向量与词向量间的相似度来决定主题-单词分布
	LF-BTM	无	无	文档分类、主题发现	吉布斯采样[67]	根据主题向量与词向量间的相似度来决定主题-单词分布
	KGE-LDA	无	无	文档分类	吉布斯采样[70]	根据主题向量与词向量间的相似度来决定

Blei 等人在 2003 年提出了第一个完全的全贝叶斯概率主题模型 LDA，由于其模型的简洁性和易扩展性，已被广泛地应用于文本挖掘、情感分析、句子分割等，然而该模型无法对主题间的相关性进行刻画以及无法捕获文本的时序特征，限制了其进一步发展；基于相关性的主题模型克服了 LDA 模型不能表达主题之间相关性的缺陷，在文本分类、新闻和社交媒体中相关性建模等领域获得了很好的应用，然而，在该类模型中，只能对成对的主题进行建模，不能实现对全局主题相关性描述；时态主题模型可以很好地对文本的时间属性进行建模，在社交媒体、在线文本、历史文献等更新较快的文本数据领域有较好的应用，但时态主题模型的建模过程通常是无监督学习过程，因此模型学习的主题可解释性低，有时往往难以理解. 此外，该类模型通常在文档生成之前，首先对其主题数目进行预先固定，这往往不切实际，因此存在一定的偏差；监督主题模型通过增加响应变量，可以有效利用文档的附加信息（Side Information），例如，评级信息和标签数据等，在处理文本分类、情感分析、标签文档聚类等问题时优于传统的 LDA 模型；结构性主题模型打破了传统的“词袋”假设，可以有效利用文档的上下文信息，进而对文档的结构信息进行分析，在词性标注、知识发现、文档结构分析等领域有较好的应用；贝叶斯非参主题模型的提出打破了主题个数是事先人为设定的假设，进而实现主题数自适应学习过程，较好地应用于主题挖掘和主题层次学习等领域，然而该类模型的后验推断是一难点所在；融合链接的主题模型可以有效地发现网络文本的潜在的结构，例如对文档的引用和被引关系建模，进而提高主题识别的准确性，在社区发现、链接发现、学术文献、网页引用等领域有较好应用，然而，该类模型随着文本数量增加，计算的复杂度会逐渐增加；情感主题模型可以从大量的文本信息中挖掘出有效的情感信息，在语义情感分类、在线用户评论等领域有较好应用；作者主题模型通过潜在的主题，将文档的作者和词汇联系起来，有效提高信息检索的准确性，主要应用于文献作者建模、文献检索等领域；词义消歧模型通常结合其背景知识或者上下文信息，来获取其正确的意思，在机器翻译、信息提取和检索以及问题问答系统等都有重要的应用；基于词向量概率主题模型均直接利用事先训练的词向量来辅助模型的学习，使得语义相近的词汇依较大概率获得同一主题，提高主题词的一致性和可解释性，丰富了文本的潜在特征表达，尤其在短文本和领域文本有较好的应用.

参数估计和推断过程是主题模型的重要组成部分，直接影响了建模的准确性与效率性. 多种概率主题模型的参数估计算法已相继提出，例如 EM 算法

(Expectation-Maximization,EM)[73]、Gibbs 采样算法[5]等.

EM 算法具有简单稳定等优势,之所以称为 EM 算法,是因为每次迭代中主要由求期望和求极大两部分组成.概率模型依赖于某些潜在变量,该算法用于求解概率模型中的参数最大似然估计,通过不断的迭代进而达到局部最优,然而该类方法不能保证全局最优.因此,使用该算法常需不断改变初始参数值或增加迭代次数来获得较优参数估计.

变分贝叶斯算法是在 EM 算法的基础上引入变分理论,也可以被认为是一种 EM 算法的扩展.该方法通过使用一个易于分解且方便优化的近似下界函数来逼近后验概率函数,进而来降低计算的复杂度,提高模型效率.然而该类方法引入的下界函数和目标函数之间存在一定的误差,导致算法精度的降低.

吉布斯采样算法是 MCMC 算法(Monte Carlo Markov-Chain algorithm)的一个特例,在每次运行的时候,首先选取概率向量的一个维度,固定其它维度变量值,再对当前的维度值进行采样,不断迭代直到满足收敛条件,最后获取待估计值.相较于 VB 算法,该算法在计算精度和算法效率上都有所提升,因此获得较为广泛的应用.然而,当文本中单词数目较大时,该类算法的抽样效率也会随之降低.

4　基于神经网络结构的主题模型

基于神经网络的主题模型主要通过神经网络来生成包含潜在主题的文本.该类模型通常将文档中的词以"词袋"形式作为输入,然后增加相应的其它网络层来生成文档,最后,利用反向求导方法对网络参数进行学习.

早期的研究主要集中于基于前馈的多层感知神经网络,之后,随着神经网络模型的快速发展,Kingma 等人提出基于变分自编码器[74]的主题模型.Card 等人考虑到在实际情况中,主题模型中分布往往具有稀疏性,因此,提出基于稀疏约束的神经主题模型[75];为了能够更好地捕获文档的上下文结构信息,Dieng 等人提出将文本单词序列作为输入的基于 RNN 结构的主题模型[76]等.

4.1　神经主题模型(Neural Topic Model,NTM)

早期,Keller 等人[77]采用多层感知器来捕获单词和文档的分布式表示,但在该模型中并不是所有层都是可解释的.随后,Cao 等人[78]提出基于前馈神经网络的主题模型①(Neural Topic Model,NTM),开始从神经网络的角度来构建主题模型.由于该模型遵循主题模型的概率特征,因此,单词和文档的分布表示具有合理的概率性解释.

在经典的 LDA 以及其扩展模型中,文档-主题分布概率矩阵表示为 $\boldsymbol{\theta}$,主题-单词分布概率矩阵表示为 $\boldsymbol{\varphi}$.在文档 d 中单词 w 的分布概率表示为 $p(w|d)=\boldsymbol{\varphi}_{.,w}\times\boldsymbol{\theta}_d^{\mathrm{T}}$.NTM 则从前馈神经网络的角度对上述两个概率分布进行描述,其中,$\boldsymbol{\varphi}_{.,w}$ 表示为带有 sigmoid 激活函数的单词查找层 $\boldsymbol{lt}$,$\boldsymbol{\theta}_d$ 表示带有 softmax 激活函数的文档查找层 $\boldsymbol{ld}$,神经网络的输出层,即文档-单词的概率分布是 $\boldsymbol{\varphi}_{.,w}$ 和 $\boldsymbol{\theta}_d$ 做点积.其相应的模型结构如图 31 所示.

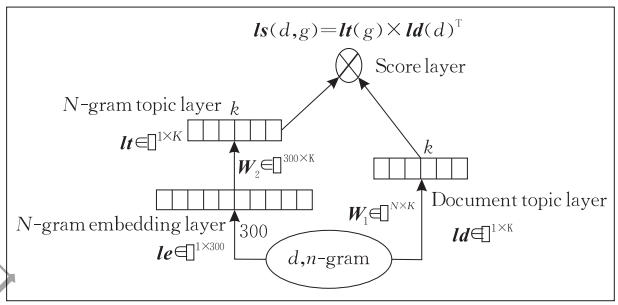

图 31　NTM 结构图

NTM 模型通过使用 sigmoid 以及 softmax 激活函数生成网络的隐藏层,如下式所示.最后,使用神经网络中常用的后向传播算法来对模型参数进行更新,进而学习出模型的两个分布以及相应的权重矩阵 $\boldsymbol{W}_1$、$\boldsymbol{W}_2$.相较于 LDA 等概率主题模型,神经主题模型不需要事先对先验分布进行假设且结构简单,但依旧可以获得较好的主题表示.

$$\boldsymbol{lt}(g)=\mathrm{sigmoid}(\boldsymbol{le}(g)\times\boldsymbol{W}_2) \tag{33}$$

$$\boldsymbol{ld}(d)=\mathrm{sofmax}(\boldsymbol{W}_1(d)) \tag{34}$$

4.2　基于变分自编码器主题模型

Miao 等人考虑使用 VAE 结构[79]来进行主题建模,进而在此基础上提出基于 VAE 结构的 NVDM 模型(Neural Variational Document Model,NVDM)[80],用于实现对文档的主题建模.NVDM 的主要思想遵从 VAE 网络结构,根据输入文档的词向量空间生成其潜在的主题特征,然后根据此潜在特征生成文档.用于文档建模的 NVDM 结构如图 32 所示.

① The implementation of NTM is available at https://github.com/elbamos/NeuralTopicModels

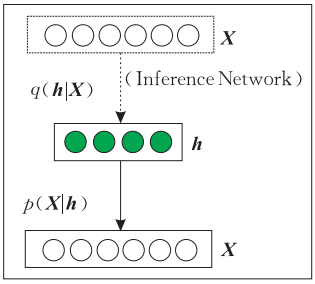

图 32 NVDM 结构

该模型将“词袋”文档表示为一个连续的潜在分布 $q(h|X)=relu(\boldsymbol{\mu}_d+\varepsilon\boldsymbol{\sigma}_d)$，其中，$\boldsymbol{\mu}_d$ 和 $\boldsymbol{\sigma}_d$ 是由 MLP 学习可得，而 ε 用于减少随机估计中的方差. 基于 softmax 激活层的解码器（生成模型）通过独立生成单词来重建文档. 值得注意的是，NVDM 仅用神经网络学习的权重矩阵 $\boldsymbol{W}$ 来描述主题-单词的分布，因此，在主题词语义一致性问题上不及 LDA.

Ding 等人[81]对上述问题进行研究，利用预训练的词向量来实现对单词对之间语义相似度的描述，相较于之前的 NVDM 模型，该方法可以显著提高了模型主题间语义的一致性.

当下大多主题模型均以文档的词汇或词向量作为输入来执行相关任务，但在实际应用中，文档还包括作者、文档来源、出版日期等元数据信息可以帮助其进行模型主题推断. Card 等人在 2018 年提出了一种将 SLDA[14]和稀疏加性生成模型（Sparse Additive Generative Models，SAGE）相结合的通用的稀疏模型框架[75]. 在该模型中，可以灵活地使用各种元数据作为标签信息来解决多标签分类问题或帮助推断预测与该标签相关的主题. 此外，该模型也可以用 SAGE 模型单位指数先验来控制主题-词汇分布的稀疏性. 由于该模型可以方便地融合元数据信息进行扩展，因此，不但可以应用于纯文本的分类或聚类，还可以灵活地应用于情感分析、时序文本数据分析等.

Gou 等人[82]考虑到动态主题模型（DTM）是对文本语料库动态表示中最流行的时间序列主题建模，然而，DTM 的后验分布需要复杂的推理过程，且建模计算时间成本高，即使是很小的变化也需要对模型进行重构，因此其可变性和通用性较差. 在此基础上，Gou 等人提出了一种新的利用变分自编码和因子图（Factor Graph，FG）来构建 DTM 的方法（VAFGDTM）. 其 VAFGDTM 网络结构如图 33 所示.

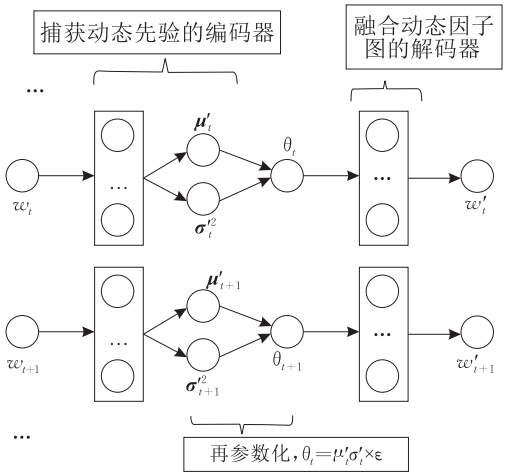

图 33 VAFGDTM 网络结构图

VAFGDTM 网络结构在时间切片 t 内，从神经网络输入层来获取文档 w_t，由编码器来计算用于生成文档分布的平均值 $\boldsymbol{\mu}'_t$ 和协方差 $\boldsymbol{\sigma}'_d$. 通过编码器学习的变分近似后验分布 $q_\phi(\theta_t|w_t)$，可以将文档 w_t 映射为文档-主题分布，与此同时，可以获取其动态先验分布，避免局部最优. 其中，生成过程中的参数 $\boldsymbol{\mu}'_t$、$\boldsymbol{\sigma}'_d$ 和 θ_t 被定义为重参数（Re-parameterize）. 解码器是用于对文档的生成概率建模的神经网络，可以将分布 $q_\phi(\theta_t|w_t)$ 映射为用于生成新文档的 w'_t 的生成概率 $p(w'_t|\theta_t,\beta_t)$，与此同时，解码器通过整合动态因子图的方式实现对时序状态变量的建模.

基于变分自编码的主题模型通常是直接利用网络隐藏层中的软最大函数来对主题模型中的假设分布进行学习，但是这种方式没有进行稀疏假设，因此，没有较好的主题提取能力. Lin 等人在此基础上，提出了一个基于 Sparse 稀疏表示的稀疏约束的神经主题模型（Neural SparseMax document and Topic Models，NSMTM）[83]. 同一般的变分自编码的主题模型类似，首先利用 MLP 学习出文档的 $\boldsymbol{\mu}_d$ 和 $\boldsymbol{\sigma}_d$，在此，使用 Sparsemax 函数[84]来产生具有稀疏表示的文档-主题和主题-单词分布，替代原来的 softmax 函数；此外，在参数学习过程中，使用 Wasserstein 散度来度量分布之间的相似度，相较与 KL 散度，可以有效增加训练的稳定性[85]；相较于 NVDM 和 AVITM，该模型在对短文本进行处理时，具有较好的泛化性能和语义一致性.

4.3 基于 RNN 结构的主题模型

上述的基于神经网络主题模型是以文档词袋的形式作为网络的输入，进而产生主题-词汇分布. 然而，在自然语言处理应用中，RNN（Recurrent Neu-

ral Network)网络结构由于可以对任意长度序列数据进行处理，生成有效的特征的优势，受到学者们的青睐.在基于 RNN 网络结构的主题训练模型中，输入层文档不再是以“词袋”形式进行输入，而是文本单词序列.输入单词序列通过 RNN 网络生成特定的潜在单元，并基于该单元生成指定主题下的自然文本.

由于 RNN 主题模型的序列性，该模型使其可以很好地捕获单词序列的局部结构，即语义和语法，但是可能在记忆长期依赖关系时遇到困难.直观地说，这些长期依赖关系具有语义性质.相反，潜在主题模型能够捕获文档的全局语义结构，但不考虑单词排序.Dieng等人在此基础上提出了 TopicRNN 模型[76]①，该模型融合了 RNNs 和潜在主题模型的优点，即它使用 RNN 捕获局部(语法)依赖关系，使用潜在主题捕获全局(语义)依赖关系，通过对两者的联合建模，提高模型在应用中的建模能力.该模型的生成过程如下所示：

对于包含词汇 $y_{1:T}$的文档：

(1)获取其主题向量 $\boldsymbol{\theta}\sim N(0,I)$；

(2)给定单词 $y_{1:t-1}$，对于在文档中的第 t 个单词 y_t：

①计算潜在单元 $h_t=f_W(x_t,h_{t-1})$，其中 $x_t\triangleq y_{t-1}$；

②获取停用词指示器 $l_t\sim\text{Bernoulli}(\sigma(\boldsymbol{\Gamma}^{\top}h_t))$，其中 σ 是 sigmoid 函数；

③生成单词 $y_t\sim p(y_t|h_t,\boldsymbol{\theta},l_t,\boldsymbol{B})$，其中 $p(y_t=i|h_t,\boldsymbol{\theta},l_t,\boldsymbol{B})\propto\exp(\boldsymbol{v}_i^{\top}h_t+(1-l_t)\boldsymbol{b}_i^{\top}\boldsymbol{\theta})$

上述停止词指示器 l_t用于控制主题向量 $\boldsymbol{\theta}$ 是否影响输出.如果 $l_t=1$ 则表示 y_t是停止词，主题向量 $\boldsymbol{\theta}$ 不影响输出；否则，将利用 $\boldsymbol{\theta}$ 与第 i 个单词的潜在词向量 $\boldsymbol{b}_i$作点积增强词汇到该主题的分配概率.因此，该模型能够对文档中出现的停用词进行自动处理，也能够使文档实现特征的自动提取.

Guo 等人为了能够更好地从文本语料库中同时捕获文档语法和全局语义关系，提出一种基于语言模型的更长上下文的循环神经网络(Larger-context Recurrent Neural Network based Language Model)[86].该模型通过动态深层主题模型提取递阶语义结构，进而指导语言文本生成.传统的基于 RNN 语言模型忽略了远程单词间的依赖和句子顺序，该模型不仅捕获了句子内部的词依赖，而且还捕获了句子之间的时间转换关系以及主题依赖.

4.4　小结

近年来，随着神经网络的逐步发展，基于神经网络的主题模型引起了研究者的关注.表 11 对比分析了各类神经网络主题模型在其网络结构和模型输入上的异同.

表 11　基于神经网络主题模型对比

模型	输入层	模型特点	网络结构	应用领域
NTM[78]	文档的n-gram向量	开始从神经网络的角度来构建主题模型，且单词和文档的分布表示具有合理的概率性解释	前馈神经网络	主题提取、文本分类
NVDM[80]	词向量	遵从 VAE 网络结构，根据输入文档词向量空间生成潜在的主题特征，然后据此潜在特征生成文档	变分自编码器	主题提取
SCHOLAR[75]	词向量	可以利用各种元数据作为标签信息来解决多标签分类问题或帮助推断预测与该标签相关的主题	变分自编码器	文本分类
VAFGDTM[82]	词袋子	通过整合动态因子图的方式实现对时序状态变量的建模	变分自编码器	信息检索、文本分类
NSMTM[83]	词向量	在基于 VAE 的主题模型建模基础上，施加稀疏约束，产生具有稀疏表示的主题和单词分布	变分自编码器	文本分类
TopicRNN[76]	文档单词序列	根据主题及上下文单词生成词汇，且可判别生成的词汇是否是停用词，能够捕获语法和语义关系	循环神经网络	单词预测、情感分析
Larger-context RNN[86]	文档单词序列	不仅捕获了句子内部的单词依赖，而且还可以捕获了句子之间的时间转换关系以及主题依赖.	循环神经网络	情感分析

但神经网络主题模型与传统概率主题模型在其模型结构和分布假设上存在较大差异.在传统概率主题模型中常需要事先对文档-主题和主题-单词分布进行假设，例如狄利克雷先验分布；而在基于神经网路的主题模型中则直接利用网络结构的结点对分布进行学习，然后使用反向传播完成模型参数训练，然而，该种分布学习方式面临最主要的问题是很难对每个维度生成的分布表示做出合理的解释.

NTM 模型直接从前馈神经网络的角度对主题模型建模中的两个分布进行描述，由于该模型未施加任何约束，模型表达较为简单直接；基于变分自编码器的主题模型在VAE基础上进一步优化主

① The implementation of TopicRNN is available at https://github.com/narratives-of-war/topic-rnn

题-单词分布假设,使其更加合理地应用于主题建模;此外,在此基础上提出的基于稀疏约束的 NSTC 模型,在其优化函数中施加主题-单词的稀疏约束,进一步提高了模型的主题语义一致性;基于 RNN 的主题模型中,输入层文档不再是以"词袋"形式输入,而是以文档的单词序列作为输入,使用 RNN 网络获取其隐藏层向量,然后根据不同学习任务输出其相应的结果,这种词序列的输入方式可以更好地捕获单词间的上下文信息,进而提高模型性能.

5 非基于 LDA 主题模型

除了当下流行的基于 LDA 概率主题模型和基于神经网络的主题模型,早期研究者们尝试在奇异值分解以及正则化等角度来对文档主题进行建模.例如,早期 Kontostathisa 等人提出了基于奇异值分解(Singular Value Decomposition,SVD)的潜在语义分析(Latent Semantic Analysis,LSA)[1]模型,在实现对文档的降维处理的同时,有效地实现对文档信息的总结提取.起初,该模型被用作一种信息检索技术,用于评估搜索引擎查询结果.后来,LSA 也被研究者们广泛应用于教育、信息系统、认知科学和人工智能等领域的研究[87-89].LSA 模型的基本思想是从一组文档中提取文本的隐含意义,首先从文档语料库中处理文本文档,也就是创建一个术语-文档矩阵,其中术语是文档中出现最频繁的单词,完成对文档术语的识别;然后,对该术语-文档矩阵进行奇异值分解,得到三个矩阵:$\boldsymbol{U}$、$\boldsymbol{S}$ 和 $\boldsymbol{V}$.其中,$\boldsymbol{U}$ 为术语特征向量矩阵,$\boldsymbol{V}$ 为文档特征向量矩阵,$\boldsymbol{S}$ 为奇异值对角矩阵.为了避免因子过拟合问题,通过保持前 k 维来截断奇异值分解矩阵的维数,完成对模型主题的提取.然而,该模型却面临"一词多义"和"多词一义"等问题.

在 LSA 模型的基础上,Hofmann 提出概率潜在语义分析(Probabilistic Latent Semantic Analysis,PLSA)模型[2],该模型是经典的 LDA 模型的前身,也称为半概率主题模型.在该模型中不考虑词序,文本语料可以由单词和文档的共现矩阵表示.从观测单词中推断两个参数:一个是将语料库中文档联系起来的全局参数,代表了给定主题后单词出现的概率;另一个是每篇文档的参数,代表了文档的主题概率分布.PLSA 模型通过引入概率统计思想,大大降低了模型的计算成本.但是,在 PLSA 模型中,对特定文档中的主题的混合比例权重没有做任何假设,因此,在实际训练时常出现过拟合情况.

对文本文档构建层次分类树是自然语言处理领域常见的一种处理方式,通过构建层次分类树能够很好地对文档间的相似性进行建模.基于此考虑,Huang 等人提出一种对文档相似性建模的 HRLSA 模型(Hierarchical -Regularized LSA)[90],该模型首先根据文档层次类结构构建文档相似图,实现所有文档类内连接,进而形成了一个优化问题,即寻找相似图中的每个节点到低维空间的最优映射,其中在该新的低维特征空间能够很好地保持原始拓扑结构中的内在联系.然后,模型将这种结构信息集成到各种学习和检索任务中,提高检索效率.然而,在实际应用中,都会因为储存容量或实现复杂度等问题,而限制模型中包含节点的个数.

为了实现对大规模数据集建模,Wang 等人提出 RLSI 模型(Regularization LSA,RLSA)[91],该模型旨在通过并行化方法将模型应用到更大文档集合中.RLSI 模型通过 l_1 或者 l_2 范数将主题模型归结为一个有正则化的二次损失函数最小化问题,这种形式化将使得模型可以凭借 MapReduce 技术[92]将学习过程分解为多个优化问题,最终实现模型并行化.这种处理方式可以方便地将模型应用于大规模数据集,然而实验证明,该种方法的效果一般.

6 主题模型的应用

随着主题模型的提出,基于主题模型的方法几乎被用到了所有的文本挖掘和智能信息处理领域,例如文本分类和聚类、信息检索、社交媒体、社区发现和图像处理等.相关主题模型使用逻辑斯蒂-正态分布,实现对主题间两两相关性进行描述,因此在社交媒体等领域有较好的应用;监督主题模型通过对每篇文档添加一个类别标识,从而进行有监督地学习,有效地应用于文本分类和情感分析等;动态主题模型通过分析文本中主题随时间演化规律,有效地应用于历史文献、用户兴趣追踪等.此外,近年来主题模型的应用已经扩展到科技文献、计算机视觉和生物信息学等.下面介绍一些基于 LDA 主题模型的代表性应用.

6.1 社交媒体

随着网络时代的到来,社交媒体,例如微博、博客等作为一种新型的媒体数据,相对于传统的文档集合具有更新速度快、内容简短等特点,具有极强的实时性[93].早期对社交媒体数据的研究,仅仅对博客领域的图结构进行分析[94].随着 LDA 主题模型的提出,Yan 等学者在此基础上提出了一种针对短

文本的 BTM(Biterm Topic Model)模型[55]，该模型是在 LDA 模型的基础上，为避免短文本中可能存在的内容稀疏问题，进而直接对短文本中的双词进行建模. 然而，BTM 模型模拟词语的共现时会引起同意文本获取不同主题的灵活性降低的问题，同时也容易引起过拟合问题. 针对社交媒体中可能出现的突发性事件问题，文献[95]提出 BBTM(Bursty Biterm Topic Model)模型，实现在微博流中突发主题建模，然而，该突发主题的生成过程是以时间作为度量标准，因此对新兴主题的识别精度较低. 动态主题模型在社交媒体中的应用，有效实现了对数据中内容和时态信息共同建模，有效分析了数据随时间的演化过程. 例如文献[11]针对动态文本流建模问题，提出一种在线 LDA 模型(On-Line LDA)，在该模型中，当有新的文本流更新的时候将增量式地更新当前模型，能够实时获取随时间变化的主题结构. 但是该模型使用离散的时间方式，因此灵活性低. 文献[96]提出一种 ETT(Emerging Topic Tracking)主题模型，从时间角度生成新兴词，而从空间角度对相关主题进行挖掘，实现对微博流中新型主题追踪；在稀疏的短文本上下文中，许多高度相关的单词可能永远不会同时出现，因此 BTM 可能会丢失许多语料库中无法观察到的、潜在的连贯和突出的词的共现模式. 为了解决这一问题，文献[97]提出了一种新的关系 BTM(R-BTM)模型，它使用词嵌入计算单词的相似列表来链接短文本；文献[98]提出了条件随机场正则化主题模型，在该模型中通过将短文本聚合成伪文档来缓解稀疏性问题，而且还利用了一个条件随机场正则化模型，使得语义相关的单词共享相同的主题分配，当应用于社交媒体短文本建模时，可以有效提高语义一致性；文献[99]在讨论线程树结构的基础上，提出一种基于流行度和传递性的会话结构感知主题模型(Conversational Structure Aware Topic Model，CSATM)来对社交媒体中在线评论进行主题推断及其评论分配；针对新闻和报道的在线社交媒体，文献[100]提出了一个新颖的基于概率主题模型的事件叙事摘要提取框架，该框架以不同的时间分辨率识别主题随时间的重复，挖掘分类时间分布，然后提取文本摘要. 该框架不仅可以从数据中捕获主题分布，还可以模拟用户活动随时间的波动，进而有效地识别主题趋势以及从带有时间戳数据的文本语料库中提取叙事摘要；针对微博短文本情感分析，文献[101]提出一种使用深度上下文文本嵌入和层析注意力机制相结合的基于方面的情感分析方法(Aspect-Based Sentiment Analysis，ABSA). 该模型是在 HAABSA 模型[102](Hybrid Approach for Aspect-Based Sentiment Analysis，HAABSA)的基础上进行改进，利用新型的基于深度上下文单词嵌入的 ELMo 模型替代传统的词向量方法，以便更好地对文本单词语义进行分析；其次，在 HAABSA 模型的基础上添加额外的注意层，使用分层注意力机制方法来进一步捕获输入数据的相关性，提高对短文本模型情感分析的能力.

6.2　图像处理

图像是一种能直观地、生动地描述客观事物的信息形式，具有较好的信息表达能力，近年来已经受到众多学者的青睐. 其中，图像分类、目标识别一直是计算机视觉研究中两个重要的问题. 特征的分析直接决定图像分类以及目标识别的准确率，进而影响人类对图像的理解. 主题模型的提出，突破了传统模型不能对图像语义进行识别的瓶颈. 文献[103]提出一种融合多特征的概率主题模型，通过 *K*-MEANS 聚类对不同特征，例如颜色、纹理、尺度等分别进行提取和量化，为语义表征提供合适的底层特征描述，使用 LDA 主题模型获取图像的语义信息. 文献[104]将监督主题模型应用于图像分类，为每幅图像添加一个全局的类别标签，将图像进行简单描述，提高图像分类准确率；文献[105]针对复杂的高维空间图像场景分类问题，提出一个完全的稀疏语义主题模型(fully sparse semantic topic model)，不但获取图像语义信息，而且可以获取主题层间场景的相关性. 而且，主题模型也应用于对图像中目标行为进行识别，例如文献[106-107]将主题模型应用于视频序列中人类行为的识别.

在图像检索中，仅仅根据图像的底层特征往往不能提取出完美的语义概念，因此，图像标注的细化已成为计算机视觉和模式识别领域的核心研究课题之一. 为了提高图像自动标注的质量，文献[108]提出了一种两阶段混合概率主题模型(two-stage hybrid probabilistic topic model)，在该模型中首先学习一个具有非对称模态的概率潜在语义分析模型来估计每个标注关键字的后验概率，在此过程中可以很好地建立图像与单词之间的关系. 然后将与相应标签相关联的图像的标签相似度和视觉相似度进行加权线性组合，构造出标签相似度图. 这样，图像底层视觉特征和高层语义概念的信息就可以通过充分考虑词与词、像与像之间的关系实现无缝集成. 最后，利用排序二松弛法进一步挖掘候选标注的相关性，从而获得细化结果，提高图像的标注精度和检索

性能;文献[109]提出了一个基于社会图像的概率主题模型,从标签和图像特征的共同出现中发现潜在主题,可以自动地将可视内容与文本标记关联起来,从而实现有效的图像搜索.

研究发现,社交媒体上的图片标签,尤其是Instagram上的图片标签,只有20%的Instagram标签描述了图片的实际内容,因此,需要应用一系列的过滤步骤来识别合适的标签.文献[110]利用LDA来预测相关图片的主题,由于主题是由一组相关术语组成的,通过所提出的方法对Instagram图像的视觉主题进行识别,进而提供了一组可信的图像标记.

在医学图像处理中,神经影像学和遗传生物标志物已被广泛地用于从鉴别的角度研究阿尔茨海默病(Alzheimer Disease,AD)的分类,文献[111]提出基于监督主题建模的AD鉴别方法,该模型中利用离散图像特征和分类遗传特征共同建模,将诊断信息-认知正常、轻度认知障碍和AD-作为监督变量引入该模型.在生成过程中引入有监督的组件可以约束模型,使其具有更强的识别性,进而提高对疾病的辨识度.

6.3 文本分类和聚类

随着网络媒体的迅猛发展,如何对海量文本进行分类,进而有效地管理和组织这些文档,成为当下重要的研究方向[112].通过对文本进行分类,用户能更加准确快速地查找到所需要的信息,方便用户对信息的浏览.文献[3]将LDA主题模型应用于文本分类,通过LDA模型将文本集表示成一个主题的概率分布,选用合适分类算法构造分类器.LDA模型对给定训练集中所有文档进行特征降维处理,有效地挖掘文本中潜在的主题信息,然而使用该种分类方法存在主题的强制分配问题;文献[113]提出一种新颖的Web网页层次分类方法,在该模型中通过使用相邻页面的附加词汇特征和主题模型进行特性表示,然后,使用基于融合矩阵的方法构造层次支持向量机的分类模型.文献[114]提出一种多标签的主题模型应用于文本分类,解决了文档只与单一类别标签相关联问题,然而,该模型忽略了多标签之间的相关性;为降低基于监督主题模型文本分类中人为标注主观性,文献[17]提出多注释的监督主题模型,通过使用多次标注降低主观性影响;文献[115]针对稀疏数据集间类的不平衡问题,提出一种基于LDA模型的重新采样的方法,使用由概率主题模型表示的类的全局语义信息来对稀疏类生成新的样本,以此解决类间不平衡的问题,提高分类准确性;针对小规模标签文档,文献[116]提出基于自我训练的半监督主题模型实现文本分类,该模型可以通过对未标记数据集的信息进行自我训练来扩大初始标签集.实验证明,该模型在小规模标签数据集上能够取得较好的实验效果,然而该模型不适用于大规模的标签文档;文献[63]提出一种新的MvTM模型,该模型从词向量空间中采样主题,并假设主题-词向量满足混合von Mises-Fisher(vMF)分布,实验的分析结果证明,该模型相较于传统的LDA模型,有效提高了主题词语义的一致性和模型的分类性能;文献[117]提出一种基于语境的深度词表示模型(Deep Contextualized word representations),该模型可以有效地捕获单词的复杂词性句法特征,也可以很好地解决同一个单词在不同语境下的不同表示问题,即单词的语义表示问题.区别于传统的为每个单词生成固定向量的词向量模型,该模型使用预训练的语言模型,首先对句子结构进行扫描,更新其内部状态,进而为句子中的每个单词都生成一个基于当前句子的词向量,也正因如此,基于深层语境的词向量表示模型,也常称为ELMo模型(Embedding for Language Models).由于该模型考虑同一单词不同语境下语义信息,因此可以有效地解决语言处理任务中的一词多义和复杂的单词语法问题,在文本分类任务上的实验结果表明,该模型可以有效提高模型分类准确性;文献[118]首次将ELMo模型应用于文本情感分类中,提出IEST模型(Implicit Emotion Shared Task),在该模型中,首先使用预训练的EMLo层来编码文本单词,然后使用一个双向长短记忆网络来丰富上下文单词表示,利用一个最大抽样步骤来对当前的单词向量创建句子表示,最终使用全连接层对句子表示进行情感分类;对上下文相关的非文字字面话语的预测分类,例如讽刺和讽刺表示,一直是自然语言中具有挑战性任务之一.文献[119]基于ELMo模型提出一种利用字符级(Character-Level)向量表示的单词模型,该模型可以捕获句子文本中复杂的形态语法特征,并将这些特征作为动态上下文中反讽或讽刺的指示符,完成对文本非字面文字话语的预测.

文本聚类是指依据同一类别中的文档尽可能相似,而不同类别中的文档间尽可能不相似的聚类假设,来实现对不同文档的聚类.文本聚类作为一种非监督的机器学习算法,不需要事先对训练样本进行标注和训练,文本聚类算法具有较好的灵活性和自动化任务处理性能,因此受到研究者的广发青睐.例如针对科技文献、热点新闻等,文本聚类可对用户感兴趣的文本进行聚类处理,将有助于用户快速浏览

和查找目标文档. 文献[120]提出一种将传统的LDA模型和K-means聚类算法相结合的一种新的文档检索模型，其中LDA模型可以有效地对文本的潜在主题语义进行分析，进而加快对用户搜索结果的反馈；文献[121]提出了一个双向量空间模型，其中语料库的每个文档都由两个向量表示：一个是基于融合的主题建模方法生成的，另一个仅仅是传统的向量模型，通过使用最先进的主题模型和数据融合方法来丰富一个集合的文档，进而有效地提高文本聚类和聚类标记的质量；文献[122]提出一种使用联合情感主题模型，将文本在低维空间进行矢量化，然后用这些向量作为文本聚类的距离度量，对不同的文章进行聚类；由于互联网上描述Web服务的文档长度较短，传统的建模方法并不理想，影响了Web服务的聚类效果. 文献[123]提出一种基于嵌入式单词和主题模型建模的方法. 首先，利用维基百科作为外部语料库对API服务文档进行扩展，再利用LF-LDA模型对其主题分布进行建模，挖掘隐含的主题信息，确定最优的主题数量，从而准确度量API服务文档之间的语义相似度，提高Web服务聚类准确性.

6.4　社区发现

社区发现通过对数据的社区结构以及演化过程进行分析，进而更好地了解网络结构的性质以及整个网络的动态趋势，从而进行网络结构优化、服务推荐等. 文献[124]通过对作者主题模型进行扩展，提出社区用户主题模型，实现了在社区发现过程中的语义建模. 文献[49]在ATM模型[47](Author Topic Model，ATM)的基础上进行扩展，提出作者接收主题模型，利用邮件的方向性确定人物在社会结构中的角色. 文献[50]将动态主题模型应用于社交网络中的社区发现，能很好地获取社交网络中的动态特征；针对如何平衡文本的网络结构问题，文献[125]提出一个基于网络正则化的统计主题模型，从而有效地对该问题进行改善；文献[126]提出一个协作主题模型，实现了对科技学术文献的社区结构发现；文献[127]提出了一种基于交互策略的LDA(interactive Latent Dirichlet Allocation)模型，在该模型中将人类专家的主观知识与LDA学习的客观知识相结合，生成意义明确的高质量主题，进而能有效地对用户生成内容进行分析，实现对学术文献研究领域结构的发现；术语出现的频率是确定文档或检索过程中某个术语重要性的常用方法，但在实际中，它通常是弱信号，特别是在频率分布平坦的情况下，比如在长关键词查询或者文本为句子、段落的短文本，术语出现的频率很低的情况下. 针对此种情况，文献[128]提出一个深度上下文术语权重框架，在该框架中，将预训练ELMo模型的上下文本表示映射为句子或段落的上下文术语权重，提高模型的术语检索和术语发现性能.

7　主题模型的数据集和评价指标及实验比较

7.1　数据集

研究者为了能够客观合理地对不同的文本主题模型在同一语料下进行对比分析，因而构造了多个比较流行的文本语料数据集，例如比较常用的20Newsgroups和Reuters等. 其中表12对当下几个常用的语料数据集名称、下载地址、文档规模以及对应的适用模型进行统计总结.

表12　主题模型常用的数据集总结

数据集名称	文档规模	应用模型
20Newsgroups①	20000	CGTM、PAM、Multi-annotator sLDA、DiscLDA、BTM
Reuters-21578②	11367	CGTM、Online-LDA、Multi-annotator sLDA
NIPS③	1740	PAM、Online-LDA、HTM、HTMM
Amazon④	10000+	SJASM、ASUM、JST、Reverse-JST
TripAdvisor⑤	10000	SJASM、MAS
Google News⑥	11109	DMM、LF-DMM
TweetSet⑦	2472	DMM、SATM、GPU-DMM

由表12可得，20Newsgroups数据集是目前应用最为广泛的语言模型评测数据集，其中包含大约20000个文本，其中，这些文本平均分布在20个不同的新闻组中，各类别的文本个数相当，因此，20Newsgroups数据集已成为许多数据挖掘任务(如文本分类)中的常用实验数据集. 此外，该数据集中每个文档都与一个类别标签相关联，对于20个新闻组，在不同的新闻组，例如(rec.sport.baseball and rec.sport.hockey)之间显示很好的相关性，因此也可用于验证主题相关性；NIPS数据集包括1988年

① The dataset is available at www.qwone.com/jason/20Newsgroups/
② The dataset is available at www.daviddlewis.com/resources/testcollections/reuters21578/
③ The dataset is available at http://www.cs.nyu.edu/~roweis/data.html
④ The dataset is available at http://www.amazon.com
⑤ The dataset is available at http://www.tripadvisor.com
⑥ The dataset is available at http://news.google.com
⑦ The dataset is available at http://trec.nist.gov/data/microblog.html

至2000年神经信息处理系统(Neural Information Processing Systems,NIPS)会议13年会议论文集,该数据集对数据进行了预处理,去掉了停止词、数字以及在语料库中出现次数少于5次的词.该数据集共包含1740篇研究论文、13 649个特殊单词和2 301 375个标记单词.值得注意的是,每篇文档都有一个时间戳,时间戳由年份决定.因此,该数据集可以很好地验证模型对文档时态性的发现;Reuters-21578数据集包含有11 367篇文档和120个标签,其中,每个文档都与多个标签关联,因此可用于对多标签的文本建模;Amazon和TripAdvisor数据集多用于对文本情感取向分析;Google News和Tweet-Set数据集用于短文本分析.

7.2 评价指标

如何对一个语言主题模型进行优劣评价一直是学者们关注的热点.目前,对主题模型好坏的评估常对模型的泛化能力、模型的复杂度、主题词语义一致性等方面.

7.2.1 模型泛化能力

困惑度(perplexity)评价标准[3]常用来对模型泛化能力进行评估,较低的困惑度表示该语言主题模型具有较好的泛化性能,建模精度较好.对于一个文本语料库,包含有N个测试文档,其中对于文档d中包含的单词个数为N_d,那么该语言模型的困惑度(Perplexity)如下所示:

$$Perplexity=-\frac{1}{N}\sum_{d=1}^{N}\frac{1}{N_d}\log P(d) \tag{35}$$

7.2.2 主题词语义一致性(Topic Coherence)

在概率主题模型建模过程中,主题词的语义是否具有较高的一致性,即是否能够生成易于理解的主题词一直是学者们所关注的问题之一.因此,为了定量地评价主题词语义一致性,Mimno等人[129]提出了主题凝聚度(coherence score),该指标能够自动地评价每个发现的主题间的一致性.主题凝聚度的主要思想是,主题词语义一致的词语在文档中往往同时出现.具体而言,给定主题z,与该主题最相关的T个单词表示为$\boldsymbol{V}^z=\{v_1^z,v_2^z,\cdots,v_T^z\}$,主题凝聚度表示为

$$C(z;\boldsymbol{V}^z)=\sum_{t=2}^{T}\sum_{l=1}^{t}\log\frac{D(v_t^z,v_l^z)+1}{D(v_l^z)} \tag{36}$$

其中,$D(v_l^z)$表示在文档中单词v_l^z的词频,$D(v_t^z,v_l^z)$表示文档中单词v_t^z和v_l^z共现的次数.但是,需要注意的是该项指标往往适合度量文档中的高频主题词,相反对于文档中出现的低频主题词度量效果较差.

另一种常用的评价方法是两两互信息(Pointwise Mutual Information,PMI)以及规一化两两互信息[40](Normalised Pointwise Mutual Information,NPMI),给定主题个数K,与主题k最相关的单词个数为T,那么PMI和NPMI计算公式如下所示:

$$PMI=\frac{1}{K}\sum_{k}\frac{2}{T(T-1)}\sum_{1\le i<j\le T}\log\frac{-p(w_i,w_j)}{p(w_i)p(w_j)} \tag{37}$$

$$NPMI=\frac{1}{K}\sum_{k}\frac{2}{T(T-1)}\sum_{1\le i<j\le T}\frac{\log\dfrac{p(w_i,w_j)}{p(w_i)p(w_j)}}{-\log p(w_i,w_j)} \tag{38}$$

其中,$p(w_i)$表示单词i出现的概率,$p(w_i,w_j)$表示单词i和j同时出现的联合概率,需要注意的是,在实验评测过程中,一般多使用大规模语料库,文档内容较长,在文档不同位置包含语义信息有一定差别,因此多使用滑动窗口来计算该联合概率.PMI和NPMI的值越高,则文档主题语义连贯性越好.

随后,Fang等人[130]提出基于词向量的主题语义一致性评测指标WESim,该评价方法通过对主题k的最相关T个单词之间语义相关程度进行计算,实现对语言主题模型的评测,经人工评测验证,该方法相较于上述两种评价方法更加适合对短文本类型的数据集的评估.

7.2.3 人工评估

对于一些文档的来源较为复杂且无标注的文本语料,使用上述评价指标效果较差,因此,在该情况下,人工评价往往更具可信性.人工评价通常是研究者直接对语言模型生成的最相关的5个或10个主题词来进行语义理解分析,进而实现对主题模型好坏的评估.

上文在不同角度详细介绍了当下语言主题模型中常用的几类流行的模型评估方法.然而在实际应用中除了上述几类评估标准,还有许多针对不同学习任务提出的评估标准,例如,在文本分类中使用准确率-召回率曲线[131]评价模型分类准确性;在聚类任务中,使用聚类纯度(Purity)和归一化互信息(Normalised Mutual Information,NMI)来对模型聚类效果进行评估[46].因此,研究者应结合实际情况来选择合适的评价指标来对模型的优劣进行合理评估.

7.3 实验分析

7.3.1 语义情感取向比较

实验选取几个经典模型进行特定刻面的语义情感取向分析(aspect-specific sentiment identification),即识别语义方面的细粒度情感.在此,使用不同模型对每篇文档的特定情感倾向分布进行推断,然而值得注意的是,对于sLDA模型,因为没有对不同层面

情绪分析模块，因此，只对评论的全局的情感进行预测，对于不同的数据集，其模型的刻面层次情感识别准确性具体如表 13 所示.

表 13　语义情感识别准确性

Model	Amazon		TripAdvisor	Average/%
	Game reviews/%	CD reviews/%	Hotel reviews/%	
LARA[115]	63.64	62.96	58.15	61.58
ASUM[45]	64.54	61.00	58.82	61.45
JST[42]	68.25	63.98	63.16	65.13
sLDA[14]	59.32	54.41	52.38	55.37
SJASM[46]	76.24	70.95	69.44	72.21

表 13 对不同模型在情绪识别准确性方面进行评估，由实验分析结果可得，SJASM 模型在 Game reviews、CD reviews 以及 Hotel reviews 三个不同数据集上都可以较好地对语义情感取向进行预测，例如在 Game reviews 数据集上准确率为 76.24%，比 sLDA、JST、ASUM 和 LARA 分别高 16.92%、7.99%、11.7%以及 12.6%.

SJASM 引入了标准线性模型，在统一框架中联合利用评审文档的总体评级作为监督数据，并为识别评审中各刻面的细粒度情感取向提供了指导和约束. 与此同时，SJASM 还利用了来自预编译的弱先验知识，对模型进行预测，因此，该模型相较于其它几类模型准确率较高. 相反，JST 和 ASUM 虽然同样利用基于情感词典的弱监督信息数据，然而在该框架中不考虑整体的评级数据证据，仅依靠文本评论内容来推断细粒度的情感，因此，影响了实验效果. LARA 在其模型结构中包含了评论的总体评分，但利用总体评分数据估计不同语义方面的权重输给 SJASM 模型进行刻面层次的情感分析. 基准模型 sLDA 利用了整体评级信息，但在建模结构中没有特定于刻面的情绪识别层. 简单地采用一个评论，将预测整体情绪作为评论中提到的各个刻面的细粒度情绪往往会导致最坏的实验结果. 例如，在现实生活中，由于一个实体各方面的效用性质不同，用户在一次评价中对某些方面的评价是正面的，而在同一次评价中对其它方面的评价则是负面的.

7.3.2　分类准确性、聚类性能和主题词语义一致性分析

随着网络社交媒体的快速发展，短文本主题建模已经引起了机器学习研究领域学者的广泛关注. 为了对模型的短文本主题建模的性能进行分析比较，本文选取几个代表性主题模型在六个不同的短文本数据集上进行实验分析，其中在每个数据集上运行每个模型 20 次，计算其平均值，然后在文本分类准确性、文本聚类性能和主题语义一致性方面进行比较分析. 其中，聚类纯度(Purity)和归一化互信息越高，则表示模型聚类性能越好；PMI 的值越高，则文档主题语义连贯性越好[132].

(1) 分类准确性分析

在对不同模型的分类性能优劣进行分析时，使用文档-主题分布来表示每篇文档，然后使用文本分类方法进行评估. 在 6 个不同数据集上，模型的分类准确性如表 14 所示. 由实验结果分析发现，模型的分类性能在一定程度上依赖于数据集. 例如不同模型在 Tweet 数据集上取得较好的分类结果，而在 PascalFlickr 数据集上效果很差. 基于词向量的 DMM 方法在分类性能上优于其它模型，尤其是在 Tweet 和 GoogleNews 数据集，这是因为 GoogleNews 和 Tweet 是通用的数据集(不是特定于某领域的)，且本文中使用的词向量是在通用数据集中训练的. 当将这些模型(LF-DMM、GPU-DMM 和 GPU-PDMM)应用于特定领域的数据集(domain-specific datasets)，模型可以通过在领域特定的数据集上重新训练词向量来进一步提高性能.

由表 14 中实验结果分析发现，基于自聚合的模型无法达到较高精度，特别是 SATM 方法，这是因为基于自聚集的方法的性能受生成伪文档的影响，如果没有任何辅助信息或元数据，生成伪文档这一步骤中的错误信息，在下一步将会被放大.

表 14　不同的文本数据集上的分类准确性比较

Model	Biomedicine	GoogleNews	PascalFlickr	Searchsnippet	StackOverflow	Tweet
LDA[3]	0.475	0.832	0.382	0.745	0.748	0.872
GSDMM[54]	0.500	0.764	0.365	0.800	0.850	0.760
LF-DMM[66]	0.443	0.832	0.372	0.755	0.756	0.858
GUP-DMM[68]	0.434	0.823	0.390	0.750	0.701	0.832
GUP-PDMM[69]	0.532	0.884	0.425	0.640	0.848	0.887
BTM[55]	0.502	0.863	0.410	0.790	0.762	0.828
WNTM[57]	0.413	0.778	0.400	0.612	0.810	0.775
SATM[58]	0.342	0.339	0.280	0.445	0.630	0.482
PTM[60]	0.442	0.825	0.380	0.700	0.730	0.875

总体来说,基于简单假设的模型,例如 BTM 和 GSDMM,在所有数据集中的分类效果优于 LDA,这表明一个文档中的两个词或所有词很可能来自一个主题.

此外,其它模型 LF-DMM、GPU-DMM、GPU-PDMM、WNTM 的性能高度依赖于数据集,例如 WNTM 模型在 Tweet、GoogleNews 和 StackOverflow 上的性能较好,但在其它数据集上的性能很差;GPU-PDMM 在除 SearchSnippets 外的所有数据集上都实现了最佳性能.

(2) 聚类性能

文本聚类是文本主题建模的一个重要应用,因此,本文将几种代表模型在不同的数据集进行聚类分析. 对于每个文档,从其主题分布中选择最大值作为聚类标签. 本文将在模型的聚类纯度和 NMI 两个方面对模型进行比较分析,如图 34 所示.

图 34 不同的文本数据集的 Purity 和 NMI

由图 34 的实验结果分析可知,除了 SATM 模型,其余几个模型的性能都优于传统的 LDA 模型. 此外同文本分类中相似性容易观察到,模型的性能高度依赖于数据集,例如 WNTM 在多个数据集上获得了最好的性能,但在 PascalFlickr 上表现很差;GPU-PDMM 在所有数据集上都表现得非常好,但在 searchsnippet 数据集上效果一般. 通过对比发现,在基于自聚合的方法中,PTM 的性能优于 SATM;对于基于全局单词出现的方法中,WNTM 在 Tweet 和 StackOverflow 数据集上的表现优于 BTM,而 BTM 在 GoogleNews 和 PascalFlickr 上的聚类效果较好;在基于 DMM 的方法中,不包含词嵌入的 GSDMM 在 Biomedicine 和 searchsnippet 数据集上聚类性能优于其它方法.

(3) 主题语义一致性

主题语义一致性是用于评价生成主题-单词分布质量好坏的重要指标,在此,仅根据单词概率为每个主题选择前 10 个单词,然后计算其 PMI 值,实验结果如表 15 所示.

表 15　不同的文本数据集上的 PMI 比较

Model	Biomedicine	GoogleNews	PascalFlickr	Searchsnippet	StackOverflow	Tweet
LDA[3]	2.112	1.064	0.896	1.061	1.125	0.988
GSDMM[54]	2.117	1.027	0.840	1.006	1.158	0.971
LF-DMM[66]	2.281	1.122	0.869	1.140	1.161	1.062
GUP-DMM[68]	2.105	1.100	0.850	1.011	1.231	1.003
GUP-PDMM[69]	2.122	1.109	0.992	0.981	1.097	1.018
BTM[55]	2.202	1.115	0.863	1.069	1.166	0.985
WNTM[57]	2.304	1.085	0.904	1.060	1.132	0.981
SATM[58]	1.981	1.029	0.851	0.814	1.136	0.829
PTM[60]	2.153	1.069	0.982	1.026	1.180	0.977

其中，LF-DMM 在 Biomedicine、GoogleNews、SearchSnippets 和 Tweet 数据集上的性能最好；GPUDMM 在 StackOverflow 上的性能最好；而 GPPDMM 在 PascalFlickr 上性能最好，这将意味着基于词向量辅助的 DMM 模型能有效缓解短文本稀疏性对建模的影响. 基于全局词共现的两种方法在每个数据集上都有很好的性能，并取得了相似的结果，这表明全局词共现的方法可以充分缓解短文本的稀疏性影响. 与上述实验结果类似，相比较而言，基于自聚合方法 PTM 和 SATM 的主题建模效果较差.

7.3.3　整合词嵌入主题模型的实验分析

近年来，基于上下文语境的深度词嵌入(Word Embeddings)向量表示模型已在自然语言处理领域产生了重要的影响. 传统的词嵌入学习表示将每个单词映射到单个矢量表示，与传统的词嵌入学习方法不同，基于上下文语境的深度词向量表示模型通常通过对语言建模来进行训练，并根据上下文对每个单词生成不同的词向量表示，进而可以有效地捕获单词的复杂词性句法特征，也可以很好地解决同一个单词在不同语境下的不同表示问题. ELMO 模型[117]是典型的基于深度上下文语境的词向量模型，该模型通过对基于多层的双向 LSTM 语言模型的不同层进行加权求和，然后使用归一化的向量作为相应的单词表示. 之后，Devlin 等人提出 BERT(Bidirectional Encoder Representations from Transformers)预训练模型[133]，该模型提出了一种新颖的 MLM(Masked Language Model)的预训练方式，其中特征抽取结构采用双向 Transformer，使得模型可以具有更深的层数，具有更好的并行性，进一步提高词向量模型地泛化能力.

JTW(Joint Topic Word-embedding)是一种基于 VAE(Variational Auto-Encoder)的语言生成模型[134]，能够同时学习文本主题以及主题的单词嵌入向量. 其中，JTW 可以很方便地与现有的基于深度语境上下文单词嵌入学习模型集成，进一步提高情绪分类等自然语言任务的性能. 具体而言，本文将 ELMo/BERT 嵌入到 JTW 中，也就是将 BOW(bag-of-words)输入替换为预先训练好的 ELMo/BERT 单词嵌入到的编码器-解码器架构中，进而使生成的单词嵌入能够更好地捕获特定领域中的语义主题. 本文在 Yelp 评论数据集中，对 JTW，ELMO 和 BERT 以及将两者整合的 JTW-ELMO 和 JTW-BERT 模型在情感分类任务上进行实验分析. 其中，对不同模型的分类结果在精度、召回、Macro-$F1$ 和 Micro-$F1$ 准则下进行比较分析，其实验结果如表 16 所示.

表 16　Yelp 数据集的情感分类性能对比

Model	Criteria			
	Precision	Recall	Macro-$F1$	Micro-$F1$
JTW	0.5713±.021	0.5639±.014	0.5599±.016	0.7339±.015
ELMO	0.6091±.005	0.6053±.001	0.6056±.002	0.7610±.005
BERT	0.6293±.014	0.5952±.006	0.6041±.012	0.7626±.005
JTW-ELMO	0.6286±.008	0.6110±.004	0.6168±.008	0.7783±.004
JTW-BERT	0.6354±.014	0.6081±.009	0.6045±.014	0.7806±.005

从表 16 的实验结果分析可得，仅使用 JTW 生成的情感分类器的结果比使用 ELMO 或 BERT 生成的分类结果要差. 然而，当将 ELMO 或 BERT 与 JTW 集成时，其组合模型(JTW-ELMO 和 JTW-BERT)的性能在 Micro-F1 评价准则计分上分别优于原始的深度上下文语境的单词向量表示模型. 其实验结果也表明，ELMO 或 BERT 等模型可以很容易地与各类语言模型进行集成，从而能够更好地获

取其特定语义信息，提高任务性能.

8 未来研究方向

8.1 智能信息处理领域的应用

当下是智能信息的时代，主题模型在智能信息处理领域，例如音乐分析、用户行为追踪、名人评论分析、图像处理等有一定的应用，但缺乏进一步深入研究，这将是未来一个重要的研究方向. 例如文献[134]重点研究了从音频信号中提取音色特征的方法，将主题模型应用于声乐音色分析；文献[135]将主题模型应用于对名人评论的情感态度分析；文献[136-137]将主题模型应用于用户行为追踪及兴趣推荐系统分析；此外，文献[138-140]分别将主题模型应用于药物的安全性评估、图像的处理、文本的可视化等，这些都将是未来重要的研究方向.

随着网络时代的到来，微博、博客、问答系统等作为新型的媒体数据，相对于传统的文档集合具有口语化、高噪声、内容简短、非规范性等特点，该类文本具有较大的研究价值，例如网络事件分析预测、文本情感分析等. 然而当下的主题模型仅在 20NewsGroups 等规范化数据集上取得了不错的建模效果，当直接对非规范文本进行建模时，模型性能将急剧下降. 因此，如何对这种开放的、非规范文本建模需进一步探索.

8.2 主题模型的扩展

对主题模型性能的扩展一直是研究的重要方向. 连续时间模型使用有向无环图并且考虑了时间和语法的依赖关系，使动态主题模型得到了有效改进. 然而，在现实世界中，要实现对文档同时考虑时间和语法的依赖关系且不受有向无环图的限制，还需进一步研究. 此外，如何利用时间戳的主题间的相关性显式地处理文档关系，也是一个重要的研究方向. 目前，已有研究基于文档显式链接，例如引用来发现文档、研究人员和社交网络之间更好的关联[141]. 例如文献[22]给出了处理文档相关性的方法，但是 Link-PLSA-LDA[73] 方法与 PLSA[2] 一样，面临着模型参数增长过快的问题，HTMM 模型(Hidden Topic Markov Model，HTMM)[22] 实现对文档结构更好建模，但不能考虑链接文档的文本内容. 因此，对以上两点进行考虑，急需一个新的模型为链接的使用提供一个有效的解决方案.

8.3 参数学习算法的扩展

参数估计和推断过程是主题模型的重要组成部分，直接影响了建模的准确性与效率性. 文献[142]提出并行的变分 EM 算法来实现对主题模型参数学习，加速训练过程；文献[143]提出平均值以及收敛吉布斯算法，提升模型的效率，然而该类算法应用于不同模型时，即便是很小的变化，仍需要重新对公式推导；文献[144]提出一个自编码器的变分贝叶斯算法，但是因为极大地内存占用，很难应用于实际场景；文献[145]提出基于黑盒推理的自编码器的变分推断算法，降低系统的内存损耗，然而通用性低. 尽管文献[146-148]也相继提出对主题模型性能的研究，然而，寻求一种高效的、灵活的、实用的参数学习算法还需进一步深入研究.

8.4 评价指标的扩展

虽然对模型优劣的衡量已经取得一些成果，但仍然还面临一些问题，例如非监督主题模型实现不同任务，例如文档分类、摘要提取[149]、信息检索[150]等，很难直接评估模型的好坏；时态主题模型在不同领域的应用将使用不同的时间切片方式，很难直接实现对模型好坏的判别.

8.5 融合高质量先验的主题模型

研究发现，基于词向量辅助的概率主题模型，可在一定程度上提高模型分类性能和主题学习能力. 然而，词向量仅仅只是对单词或者潜在概念间的语义相似程度间的距离进行度量，无法实现对单词或概念间的关系进行推理或表示. 其中，基于表示学习的知识图谱不但能够完成对实体间关系的表示，还可实现对实体链接的预测和推理. Yao 等人已开始尝试利用知识图嵌入来提升主题模型的建模能力. 因此，能否利用当下知识图谱已有的研究成果在主题建模中融合高质量的先验知识来进一步提高模型的表达能力，将是未来一个重要的研究问题.

9 结 论

概率主题模型是自然语言处理领域的一个重要的研究方向，由于其出色的对高维数据的层次结构化建模能力受到研究者的广泛青睐. 本文系统地对各种层次主题模型进行了综述，指出了各个模型提出的原因、所具有的优缺点以及模型的典型应用.

纵观概率主题模型的发展历程，自 2003 年 LDA 主题模型被提出以来便受到了广泛应用，针对其在应用过程中的不足，文章在第 2 节首先对其基本数学概念进行介绍，然后在第 3 节详细地对基于 LDA 模型的各类扩展模型的生成过程以及优缺点进行详细介绍，例如相关主题模型、时态主题模型、监督主题模型以及非参数模型. 此外，近年来，为提

高模型的实用性，面向特定任务的主题模型的研究，例如词义消歧、链接发现、情感分析、作者主题模型等相继出现；在文章的第 4 节对近年来提出的几类经典的基于神经网络的主题模型进行介绍；在文章的第 5 节对非基于 LDA 主题模型进行介绍；在文章的第 6 节，主要介绍了主题模型近年来在文本分类、文本聚类、社交媒体、图像处理以及社区发现等领域的主要应用. 在文章的第 7 节，主要介绍了概率主题模型常用的几个公认的数据集、评测方法以及典型实验结果.

最后，在文章的第 8 节详细阐述了主题模型在未来的研究方向，进一步说明概率主题模型的巨大的应用潜力.

总之，主题模型现已被广泛应用于各个领域，例如计算机视觉、生物信息学、文本挖掘等，由于其独特的对大规模数据降维的优势，主题模型必将发挥越来越重要的作用.

参 考 文 献

[1] Kontostathisa A, Pottengerb W M. A framework for understanding latent semantic indexing (LSI) performance. Information Processing & Management, 2006, 42(1): 56-73

[2] Hofmann T. Probabilistic latent semantic analysis//Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI). Stockholm, Sweden, 1999: 123-145

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

[4] Blei D M, Lafferty J D. A correlated topic model of science. The Annals of Applied Statistics, 2007, 1(1): 17-35

[5] Li W, Jordan N. Pachinko allocation: DAG-structured mixture models of topic correlations//Proceedings of the International Conference on Machine Learning. Pittsburgh, USA, 2006: 577-584

[6] Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model. Journal of Machine Learning Research, 2003, 3(2): 1137-1155

[7] Xun Guangxu, Li Yaliang. A correlated topic model using word embeddings//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 4207-4213

[8] Hu Pengfei, Liu Wenju, Jiang Wei, Yang Zhanlei. Latent topic model based on Gaussian-LDA for audio retrieval//Proceedings of the Chinese Conference on Pattern Recognition. Springer, 2012: 556-563

[9] Suzanna Sia, Ayush Dalmia, Sabrina J. Mielke. Tired of topic models? Clusters of pretrained word embeddings make for fast and good topics too !//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Online, 2020: 1728-1736

[10] Blei D M, Lafferty J D. Dynamic topic models//Proceedings of the International Conference on Machine Learning. Pittsburgh, USA, 2006: 113-120

[11] Alsumait L, Barbará D, Domeniconi C. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking//Proceedings of the 8th IEEE International Conference on Data Mining. Pisa, Italy, 2008: 3-12

[12] Wang C, Blei D, Heckerman D. Continuous time dynamic topic models//Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence. Helsinki, Finland, 2008: 579-586

[13] Erics K. An introduction to probability and stochastic processes. The Society for Industrial and Applied Mathematics, 1994, 36(1): 128-129

[14] Blei D M, McAuliffe J. Supervised topic models//Proceedings of the Neural Information Processing Systems. Vancouver, Canada, 2007: 121-128

[15] Lacoste-Julien S, Jordan M I. DiscLDA: Discriminative learning for dimensionality reduction and classification//Proceedings of the Neural Information Processing Systems. Vancouver, Canada, 2007: 897-904

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

[17] Rodrigues F, Lourenco M, Ribeiro B. Learning supervised topic models for classification and regression from crowds. IEEE Transactions on Pattern Analysis and Machine Intelligenc, 2017, 39(12): 2409-2422

[18] Mei Q, Shen X, Zhai C X. Automatic labeling of multinomial topic models//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA, 2007: 490-499

[19] Lau J H, Grieser K, Newman D, Baldwin T. Automatic labelling of topic models//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, USA, 2011: 1536-1545

[20] Wadsworth W D, Argiento R, Guindani M. An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. BMC Bioinformatics, 2017, 18(1): 185-189

[21] Zhao H, Du L, Buntine W, et al. MetaLDA: A topic model that efficiently incorporates meta information//Proceedings of the IEEE International Conference on Data Mining (ICDM). New Orleans, LA, USA, 2017: 635-644

[22] Griffiths T L, Steyvers M, Blei D M. Integrating topics and syntax//Proceedings of the Neural Information Processing Systems. Vancouver, Canada, 2004: 537-544

[23] Gruber A, Rosen-Zvi M, Weiss Y. Hidden topic Markov models//Proceedings of the Artificial Intelligence and Statistics (AISTATS). San Juan, USA, 2007: 21-24

[24] Mccallum A, Freitag D, Pereira F. Maximum entropy markov models for information extraction and segmentation//Proceedings of the International Conference on Machine Learning. California, USA, 2000: 1-26

[25] Charles A S, Andrew M, Khashayar R. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Journal of Machine Learning Research, 2007, 8(4): 693-723

[26] Teh Y W, Jordan M I. Hierarchical dirichlet processes. Journal of the American Statiscal Association, 2006, 101(476): 1566-1581

[27] Blei D M, Jordan M I, Griffiths T L. Hierarchical topic models and the nested Chinese restaurant process//Proceedings of the Neural Information Processing Systems. Vancouver, Canada, 2003: 17-24

[28] Basili R, Giannone C, Croce D. Latent topic models of surface syntactic information//Proceedings of the Artificial Intelligence Around Man and Beyond. Palermo, Italy, 2011: 225-237

[29] Lu J, Xuan J, Zhang G. Bayesian nonparametric relational topic model through dependent Gamma processes. IEEE Transactions on Knowledge and Data Engineering, 2016, 29(7): 1357-1369

[30] Lim K W, Buntine W, Chen C. Nonparametric Bayesian topic modelling with the hierarchical Pitman-Yor processes. International Journal of Approximate Reasoning, 2016, 78(2): 172-191

[31] Cohn D, Hofmann T. The missing link—A probabilistic model of document content and hypertext connectivity//Proceedings of the International Conference on Neural Information Processing Systems. Denver, USA, 2000: 430-436

[32] Cohn D, Chang H. Learning to probabilistically identify authoritative documents//Proceedings of the 17th International Conference on Machine Learning. Stanford, USA, 2000: 167-174

[33] Erosheva E, Fienberg S, Lafferty J. Mixed-membership models of scientific publications. Proceedings of the National Academy of Sciences, 2004, 101(1): 5220-5227

[34] Nallapati R M, Ahmed A, Xing E P. Joint latent topic models for text and citations//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 542-550

[35] Airoldi E M, Blei D M, Fienberg S E, Xing E P. Mixed membership stochastic block models for relational data, with applications to protein interactions//Proceedings of the 8th IEEE International Conference on Data Mining. Pisa, Italy, 2006: 1045-1055

[36] Zhang A, Zhu J, Zhang B. Sparse relational topic models for document networks//Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. New York, USA, 2013: 670-685

[37] Taskar B, Abbeel P, Koller D. Discriminative probabilistic models for relational data//Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence. Edmonton, Canada, 2002: 485-492

[38] Wang Hao, Shi Xingjian, Yeung Dit-Yan. Relational deep learning: A deep latent variable model for link prediction//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 2688-2094

[39] Terragni S, Fersini E, Messina E. Constrained relational topic models. Information Sciences, 2020, 512(1): 581-594

[40] Titov I, Mcdonald R. Modeling Online Reviews with Multi-grain Topic Models//Proceedings of the 17th International Conference on World Wide Web. Beijing, China, 2008: 111-120

[41] Zhu J, Zhu M, Wang H. Aspect-based sentence segmentation for sentiment summarization//Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion. New York, USA, 2009: 65-72

[42] Williamson C L, Zurko M E, Patel-Schneider P F, Shenoy P J. Topic sentiment mixture: Modeling facets and opinions in weblogs//Proceedings of the 16th International Conference on World Wide Web. Banff, Canada, 2007: 171-180

[43] Lin C, He Y. Joint sentiment/topic model for sentiment analysis//Proceedings of the 18th ACM Conference on Information and Knowledge Management. Hong Kong, China, 2009: 375-384

[44] Lin C, He Y, Everson R. Weakly supervised joint sentiment-topic detection from text. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(6): 1134-1145

[45] Jo Y, Oh A H. Aspect and sentiment unification model for online review analysis//Proceedings of the 4th International Conference on Web Search and Web Data Mining. Hong Kong, China, 2011: 815-824

[46] Hai Z, Cong G, Chang K. Analyzing sentiments in one go: A supervised joint topic modeling approach. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(6): 1172-1185

[47] Steyvers M, Smyth P, Rosen-Zvi M, Griffiths T L. Probabilistic author-topic models for information discovery//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, USA, 2004: 306-315

[48] Tang J, Jin R, Zhang J. A topic modeling approach and its integration into the random walk framework for academic search//Proceedings of the 8th IEEE International Conference on Data Mining. Pisa, Italy, 2008: 1055-1060

[49] 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. Computer Science Department Faculty Publication Series, 2005, 44(1): 1-17

[50] Xu S, Shi Q, Qiao X, et al. Author-Topic over Time (AToT): A dynamic users' interest model//Proceedings of the 4th International Conference on Mobile, Ubiquitous, and Intelligent Computing. Gwangju, Korea, 2013: 239-245

[51] Boyd-Graber J, Blei D. A topic model for word sense disambiguation//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, Czech Republic, 2007: 1024-1033

[52] Chaplot D S, Salakhutdinov R. Knowledge-based word sense disambiguation using topic models//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 5062-5069

[53] McCarthy D, Koeling R, Weeds J, Carroll J A. Finding predominant word senses in untagged text//Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain, 2004: 279-286

[54] Zhao Wayne Xin, Jiang Jing, Weng Jianshu, He Jing. Comparing twitter and traditional media using topic models//Proceedings of the Advances in Information Retrieval-33rd European Conference on IR Research. Dublin, Ireland, 2011: 338-349

[55] Yan Xiaohui, Guo Jiafeng, Lan Yanyan, Cheng Xueqi. A biterm topic model for short texts//Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil, 2013: 1445-1456

[56] Wang W, Zhou H, He K. Learning latent topics from the word co-occurrence network//Proceedings of the 35th National Conference of the Theoretical Computer Science. Wuhan, China, 2017: 18-30

[57] Zuo Y, Zhao J, Xu K. Word network topic model: A simple but general solution for short and imbalanced texts. Knowledge and Information Systems, 2016, 48(2): 379-398

[58] Quan X, Kit C, Ge Y, et al. Short and sparse text topic modeling via self-aggregation//Proceedings of the International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 2270-2276

[59] Shi Lei, Du Junping, Liang Meiyu, Kou Feifei. Dynamic topic modeling via self-aggregation for short text streams. Peer-to-Peer Networking and Applications, 2019, 12(5): 1403-1417

[60] Zuo Y, Wu J, Zhang H. Topic modeling of short texts: A pseudo-document view//Proceedings of the International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 2105-2114

[61] Yu Ke-Ren, Fu Yun-Bin, Dong Qi-Wen. Survey on distributed word embeddings based on neural network language models. Journal of East China Normal University (Natural Science), 2017, (5): 52-65(in Chinese)
(郁可人，傅云斌，董启文. 基于神经网络语言模型的分布式词向量研究进展. 华东师范大学学报(自然科学版)，2017，(5)：52-65)

[62] Das R, Zaheer M, Dyer C. Gaussian LDA for topic models with word embeddings//Proceedings of the Annual Meeting of the Association for Computational Linguistics and the Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Beijing, China, 2015: 795-804

[63] Li X, Chi J, Li C, et al. Integrating topic modeling with word embeddings by mixtures of vMFs//Proceedings of the International Conference on Computational Linguistics. Osaka, Japan, 2016: 151-160

[64] Gopal S, Yang Y. von Mises-Fisher clustering models//Proceedings of the International Conference on Machine Learning. Beijing, China, 2014: 154-162

[65] He J, Hu Z, Bergkirkpatrick T, et al. Efficient correlated topic modeling with topic embedding//Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Halifax, Canada, 2017: 225-233

[66] Nguyen D Q, Billingsley R, Du L, Johnson M. Improving topic models with latent feature word representations. Transactions of the Association for Computational Linguistics, 2015, (3): 299-313

[67] Liu Liang-Xuan, Huang Meng-Xing. Biterm topic model with word vector features. Application Research of Computers, 2017, 34(7): 2055-2058(in Chinese)
(刘良选，黄梦醒. 融合词向量特征的双词主题模型. 计算机应用研究，2017，34(7)：2055-2058)

[68] Li C, Wang H, Zhang Z, et al. Topic modeling for short texts with auxiliary word embeddings//Proceedings of the International ACM SIGIR Conference on Research & Development in Information Retrieval. Pisa, Italy, 2016: 165-174

[69] Li C, Duan Y, Wang H, et al. Enhancing topic modeling for short texts with auxiliary word embeddings. ACM Transactions on Information Systems, 2017, 36(2): 1-30

[70] Yao L, Zhang Y, Wei B, et al. Incorporating knowledge graph embeddings into topic modeling//Proceedings of the National Conference on Artificial Intelligence. San Francisco, USA, 2017: 3119-3126

[71] Bordes A, Usunier N, Garcia-Duran A. Translating embeddings for modeling multi-relational data//Proceedings of the National Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2013: 2787-2795

[72] Li D, Dadaneh S Z, Zhang J, Li P. Integration of knowledge graph embedding into topic modeling with hierarchical Dirichlet process//Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL), Minneapolis, USA, 2019: 940-950

[73] Xu Ge, Wang Hou-Feng. Development of topic models in natural language processing. 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. 计算机学报，2011，34(8)：1423-1436)

[74] Kingma D P, Welling M. Auto-encoding variational Bayes//Proceedings of the International Conference on Learning Representations(ICLR). Banff, Canada, 2014: 12-14

[75] Card D, Tan C, Smith N A, et al. Neural models for documents with metadata//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia, 2018: 2031-2040

[76] Dieng A B, Wang C, Gao J, et al. TopicRNN: A recurrent neural network with long-range semantic dependency//Proceedings of the International Conference on Learning Representations. Toulon, France, 2017: 1-8

[77] Keller M, Bengio S. A neural network for text representation //Proceedings of the Artificial Neural Networks: Formal Models and Their Applications(ICANN). Warsaw, Poland, 2005: 667-672

[78] Cao Z, Li S, Liu Y, et al. A novel neural topic model and its supervised extension//Proceedings of the National Conference on Artificial Intelligence. Austin, USA, 2015: 2210-2216

[79] Romain L, Jeffrey R, Michael I J. Information constraints on auto-encoding variational Bayes//Proceedings of the International Conference on Neural Information Processing Systems. Montréal, Canada, 2018: 6117-6128

[80] Miao Y, Yu L, Blunsom P. Neural variational inference for text processing//Proceedings of the International Conference on Machine Learning. New York, USA, 2016: 1727-1736

[81] Ding R, Nallapati R, Xiang B. Coherence-aware neural topic modeling//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 830-836

[82] Gou Zhinan, Han Lixin, Sun Ling, et al. Constructing dynamic topic models based on variational autoencoder and factor graph. IEEE Access, 2018, 6(1): 53102-53111

[83] Lin T, Hu Z, Guo X, et al. Sparsemax and relaxed Wasserstein for topic sparsity//Proceedings of the ACM International Conference on Web Search and Data Mining. Melbourne, Australia, 2019: 141-149

[84] Martins A F, Astudillo R F. From softmax to sparsemax: A sparse model of attention and multi-label classification//Proceedings of the International Conference on Machine Learning. New York, USA, 2016: 1614-1623

[85] Huan Jia-Jia, Li Peng-Wei, Peng Min, et al. Review of deep learning-based topic model. Chinese Journal of Computers, 2020, 43(5): 827-855(in Chinese)
(黄佳佳, 李鹏伟, 彭敏等. 基于深度学习的主题模型研究. 计算机学报, 2020, 43(5): 827-855)

[86] Guo D, Chen B, Lu R. Recurrent hierarchical topic-guided neural language models//Proceedings of the 37th International Conference on Machine Learning. Vienna, Australia, 2020: 3810-3821

[87] Han J, Kamber M. Data Mining: Concepts and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems). Netherlands: Elsevier, 2006

[88] Mitra B, Craswell N. An introduction to neural information retrieval. Foundations and Trends in Information Retrieval, 2018, 13(1): 1-126

[89] Landauer T K. LSA as a theory of meaning//Landauer T, McNamara D S, Dennis S, Kintsch W, eds. Handbook of Latent Semantic Analysis. Washington, USA: Lawrence Erlbaum Associates Publishers, 2007

[90] Huang Yi, Yu Kai, Schubert M. Hierarchy regularized latent semantic indexing//Proceedings of the Data Mining. Houston, USA, 2005: 178-185

[91] Wang Quan, Xu Jun, Li Hang, Craswell N. Regularized latent semantic indexing//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 685-694

[92] Guo Xiaoguang, Guo Zhigao, Ren Hao. Learning Bayesian network parameters via minimax algorithm. International Journal of Approximate Reasoning, 2019, 108(1): 62-75

[93] Chai Bian-Fang, Jia Cai-Yan, Yu Jian, et al. Survey of probabilistic models combining content and link for network structure detection. Journal of Chinese Computer Systems, 2013, 34(11): 2524-2528(in Chinese)
(柴变芳, 贾彩燕, 于剑等. 融合内容和链接的网络结构发现概率模型综述. 小型微型计算机系统, 2013, 34(11): 2524-2528)

[94] Kumar R, Novak J, Raghavan P. On the bursty evolution of blogspace. World Wide Web, 2005, 8(2): 159-178

[95] Yan X, Guo J, Lan Y. A probabilistic model for bursty topic discovery in microblogs//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, Texas, USA, 2015: 353-359

[96] Huang J, Peng M, Wang H. A probabilistic method for emerging topic tracking in microblog stream. World Wide Web, 2016, 20(2): 23-33

[97] Li Ximing, Zhang Ang, Li Changchun, et al. Relational biterm topic model: Short-text topic modeling using word embeddings. Computer Journal, 2019, 62(3): 359-372

[98] Gao W, Peng M, Wang H. Incorporating word embeddings into topic modeling of short text. Knowledge and Information Systems, 2019, 61(2): 1123-1145

[99] Sun Y, Loparo K, Kolacinski R. Conversational structure aware and context sensitive topic model for online discussions//Proceedings of the 2020 IEEE 14th International Conference on Semantic Computing (ICSC), San Diego, USA, 2020: 85-92

[100] Oghaz T A, Mutlu E C, Jasser J, et al. Probabilistic model of narratives over topical trends in social media, a discrete time model//Proceedings of the 31st ACM Conference on Hypertext and Social Media. Virtual Event, USA, 2020: 281-290

[101] Trusca M M, Wassenberg D, Frasincar F, Dekker R. A hybrid approach for aspect-based sentiment analysis using deep contextual word embeddings and hierarchical attention//Proceedings of the International Conference on Web Engineering. Helsinki, Finland, 2020: 365-380

[102] Wallaart O, Frasincar F. A hybrid approach for aspect-based sentiment analysis using a lexicalized domain ontology and attentional neural models//Proceedings of the Extended Semantic Web Conference. Portorož, Slovenia, 2019: 363-378

[103] Zhong Y, Zhu Q, Zhang L. Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(11): 1-16

[104] Wang C W C, Blei D, Li F F. Simultaneous image classification and annotation//Proceedings of the Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 1903-1910

[105] Zhu Q, Zhong Y, Zhang L. Scene classification based on the fully sparse semantic topic model. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(10): 5525-5538

[106] Wang Y, Mori G. Human action recognition by semi-latent topic models//Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence. Miami, USA, 2009: 1762-1774

[107] Hospedales T M, Gong Shaogang, Xiang Tao. A Markov clustering topic model for mining behaviour in video//Proceedings of the 12th International Conference on Computer Vision. Kyoto, Japan, 2009: 1165-1172

[108] Tian D, Shi Z. A two-stage hybrid probabilistic topic model for refining image annotation. International Journal of Machine Learning & Cybernetics, 2020, 11(2): 417-431

[109] Tu N A, Khan K U, Lee Y-K. Featured correspondence topic model for semantic search on social image collections. Expert Systems with Applications, 2018, 77(1): 20-33

[110] Argyrou A, Giannoulakis S, Tsapatsoulis N. Topic modelling on Instagram hashtags: An alternative way to automatic image annotation?//Proceedings of the 2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP). Zaragoza, Spain, 2018: 61-67

[111] Yang J, Feng X, Laine A F, Angelini E D. Characterizing Alzheimer's disease with image and genetic biomarkers using supervised topic models. IEEE Journal of Biomedical and Health Informatics, 2020, 24(4): 1180-1187

[112] Gui Xiao-Qing, Zhang Jun, Zhang Xiao-Min. Survey on temporal topic model methods and application. Computer Science, 2017, 44(2): 46-55(in Chinese)
(桂小庆，张俊，张晓民. 时态主题模型方法及应用研究综述. 计算机科学，2017，44(2)：46-55)

[113] Deng Li, Du Xin, Shen Jizhong. Web page classification based on heterogeneous features and a combination of multiple classifiers. Frontiers of Information Technology & Electronic Engineering, 2020, 21(7): 995-1004

[114] Rubin T N, Chambers A, Smyth P, Steyvers M. Statistical topic models for multi-label document classification. Machine Learning, 2012, 88(1): 157-208

[115] Chen E, Lin Y, Xiong H. Exploiting probabilistic topic models to improve text categorization under class imbalance. Information Processing & Management, 2011, 47(2): 202-214

[116] Pavlinek M, Podgorelec V. Text classification method based on self-training and LDA topic models. Expert Systems with Applications, 2017, 80(1): 83-93

[117] Peters M E, Neumann M, Iyyer M. Deep contextualized word representations//Proceedings of the North American Chapter of the Association for Computational Linguistics. New Orleans, USA, 2018: 2227-2237

[118] Balazs J A, Marrese-Taylor E, Matsuo Y. Implicit emotion classification with deep contextualized word representations//Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Brussels, Belgium, 2018: 50-56

[119] Ilic S, Marrese-Taylor E, Balazs J A, Matsuo Y. Deep contextualized word representations for detecting sarcasm and irony//Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Brussels, Belgium, 2018: 2-7

[120] Ruan Guang-Ce, Xia Lei. Research on retrieval result clustering based on topic model. Intelligence Magazine, 2017, 36(3): 179-184(in Chinese)
(阮光册，夏磊. 基于主题模型的检索结果聚类应用研究. 情报杂志，2017，36(3)：179-184)

[121] Pourvali M, Orlando S, Omidvarborna H. Topic models and fusion methods: A union to improve text clustering and cluster labeling. International Journal of Interactive Multimedia and Artificial Intelligence, 2019, 5(4): 28-34

[122] Sánchez O, Sierra G. Joint sentiment topic model for objective text clustering. Journal of Intelligent and Fuzzy Systems, 2019, 36(4): 3119-3128

[123] Chen Yanping, Wang Xin, Xia Hong, et al. Research on web service clustering method based on word embedding and topic model//Proceedings of the International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. Kunming, China, 2019: 980-987

[124] Zhou D, Manavoglu E, Li J. Probabilistic models for discovering e-communities//Proceedings of the 15th International Conference on World Wide Web. Edinburgh, UK, 2006: 173-182

[125] Mei Q, Cai D, Zhang D. Topic modeling with network regularization//Proceedings of the 17th International Conference on World Wide Web. Beijing, China, 2008: 101-110

[126] Wang C, Blei D M. Collaborative topic modeling for recommending scientific articles//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 448-456

[127] Liu Yezheng, Du Fei, Sun Jianshan. iLDA: An interactive latent Dirichlet allocation model to improve topic quality. Journal of Information Science, 2020, 46(1): 1-8

[128] Dai Zhuyun, Callan J. Context-aware sentence/passage term importance estimation for first stage retrieval// Proceedings of the Association for Computing Machinery. New York, USA, 2020: 1533-1536

[129] Mimno D M, Wallach H M, Talley E M, et al. Optimizing semantic coherence in topic model//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK, 2011: 262-272

[130] Fang A, Macdonald C, Ounis I, et al. Using word embedding to evaluate the coherence of topics from Twitter data// Proceedings of the 39th International ACM SIGIR Conference. Pisa, Italy, 2016: 1057-1060

[131] Jordan M I, Ghahramani Z, Jaakkola T, Saul L. An introduction to variational methods for graphical models. Machine Learning, 1999, 37(2): 183-233

[132] Qiang Jipeng, Qian Zhenyu, Li Yun, Yuan Yunhao. Short text topic modeling techniques, applications, and performance: A survey. Computing Research Repository, 2019, 14(8): 1-17

[133] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, Minnesota, 2019: 4171-4186

[134] Nakano T, Yoshii K, Goto M. Vocal timbre analysis using latent Dirichlet allocation and cross-gender vocal timbre similarity//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy, 2014: 5202-5206

[135] Shi B, Lam W, Bing L, Xu Y. Detecting common discussion topics across culture from news reader comments//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, 2016: 1104-1110

[136] Giri R, Choi H, Hoo K S, Rao B. User behavior modeling in a cellular network using latent Dirichlet allocation// Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning. Springer, Germany, 2014: 36-44

[137] Wang Shuhui, Wang Zhenjun, Jiang Shuqiang. Cross media topic analytics based on synergetic content and user behavior modeling//Proceedings of the Multimedia and Expo (ICME). Chengdu, China, 2014: 1-6

[138] Bui T X, Sprague Jr R H. Extracting consumer health expressions of drug safety from web forum//Proceedings of the 48th Hawaii International Conference on System Sciences. Hawaii, USA, 2015: 2896-2905

[139] Vaduva C, Gavat I, Datcu M. Latent Dirichlet allocation for spatial analysis of satellite images. IEEE Transactions on Geoscience and Remote Sensing, 2013, 51(5): 2770-2786

[140] Siersdorfer S, Chelaru S, Pedro J S, et al. Analyzing and mining comments and comment ratings on the social web. Association for Computing Machinery (ACM) Transactions on the Web, 2014, 8(3): 1-39

[141] Azarbonyad H, Dehghani M, Kenter T. HiTR: Hierarchical topic model re-estimation for measuring topical diversity of documents. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(11): 2124-2137

[142] Nallapati R, Cohen W W, Lafferty J D. Parallelized variational EM for latent Dirichlet allocation: An experimental evaluation of speed and scalability//Proceedings of the 7th IEEE International Conference on Data Mining. Omaha, USA, 2007: 349-354

[143] Griffiths T L, Steyvers M. Finding scientific topics. The National Academy of Sciences, 2004, 101(1): 5228-5235

[144] Newman D, Lau J H, Grieser K, Baldwin T. Automatic evaluation of topic coherence//Proceedings of the Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics. Los Angeles, USA, 2010: 100-108

[145] Srivastava A, Sutton C. Autoencoding variational inference for topic models//Proceedings of the International Conference on Learning Representations (ICLR). Toulon, France, 2017: 23-28

[146] Porteous I, Newman D, Ihler A, Asuncion A. Fast collapsed Gibbs Sampling for lantent Dirichlet allocation//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 569-577

[147] Anandkumar A, Foster D P, Hsu D. A spectral algorithm for latent Dirichlet allocation. Algorithmica, 2015, 72(1): 193-214

[148] Chien J T, Lee C H. Deep unfolding for topic models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(2): 318-331

[149] Petinot Y, Mckeown K, Thadani K. A hierarchical model of web summaries//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, USA, 2011: 670-675

[150] Wei X, Croft B. LDA-based document models for ad-hoc retrieval//Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, USA, 2006: 178-185

HAN Ya-Nan, Ph. D. candidate. Her main research interest is machine learning.

LIU Jian-Wei, Ph. D., associate professor. His main research interests include machine learning, pattern recognition and intelligent system, analysis, prediction and control of complex system, algorithm analysis and design.

LUO Xiong-Lin, Ph. D., professor. His main research interests include intelligent control, and analysis, prediction, controlling of complicated nonlinear system.

Background

Topic model is one of the most important techniques in text mining, which is widely used in data mining, text classification and community discovery. Topic model has become a hot direction in the field of natural language processing because of the excellent dimensionality reduction ability and the flexible and extensible ability to construct probabilistic model.

This paper systematically introduces the LDA model, makes a particular categorization on topic models derived from LDA, and then points the motivation of every topic model, the advantages of every topic model, the problems that every topic model can solve, the form of every topic model, and the typical application scenarios that topic models can be used. In addition, several commonly datasets, evaluation metrics and typical experimental results of probability topic models are introduced in detail. Finally, we reveal the problems and the research directions of the probabilistic topic models in the future.

This work is supported by the Science Foundation of China University of Petroleum, Beijing (No. 2462020YXZZ023).