文章编号:1007－6735(2019)03－0273－08　DOI: 10.13255/j.cnki.jusst.2019.03.010

基于 LDA 模型的网络刊物主题发现与聚类

杨传春[1]，　张冰雪[2]，　李仁德[1]，　郭　强[1]
（1. 上海理工大学 复杂系统科学研究中心，上海 200093；2. 上海理工大学 MPA 教育中心，上海 200093）

摘要： 随着智能终端的普及，文本的主题挖掘需求也越来越广泛，主题建模是文本主题挖掘的核心，LDA 生成模型是基于贝叶斯框架的概率模型，它以语义关联为基础，很好地解决了文本潜在主题的提取问题。对文本聚类过程的核心技术 LDA 生成模型、数据采样、模型评价等作了较为深入的阐述和解析，结合网络教育平台的 2 794 篇学习刊物进行了主题发现和聚类实验，建立了包含 3 800 个词项的词库，通过 kmeans 算法和合并向量算法（UVM）分两步解决了主题聚类问题。提出了文本挖掘实验的一般方法，并对层次聚类中文本距离的算法提出了改进。实验结果表明，该平台刊物的主题整体相似度比较好，但主题过于集中使得许多刊物的内容不具有辨识度，影响用户对主题的定位。

关键词： LDA 模型；生成模型；主题发现；层次聚类；文本挖掘
中图分类号： N 32　　**文献标志码：** A

Topic Discovery and Clustering for Online Journals Based on LDA Algorithm

YANG Chuanchun[1], ZHANG Bingxue[2], LI Rende[1], GUO Qiang[1]
(1. *Research Center of Complex Systems Science, University of Shanghai for Science and Technology, Shanghai 200093, China*; 2. *MPA Education Center, University of Shanghai for Science and Technology, Shanghai 200093, China*)

Abstract: With the popularity of intelligent terminals, the demand of text topic mining is becoming more prevalent in many different domains. Theme modeling is the kernel of text topic mining. LDA (latent Dirichlet allocation) generating model is a probability model based on Bayesian framework, and it solves the problem of text potential topic extraction based on semantic association. The key technology of text clustering process, including LDA generating model, data sampling, model evaluation, was described and analyzed in depth. Theme discovery and clustering experiments were carried out in 2 794 learning journals on the network education platform. A thesaurus containing 3 800 terms was established. The problem of topic clustering was solved by kmeans algorithm and UVM (union vector method) algorithm in two steps. Meanwhile a general method of text mining experiment was proposed, and the algorithm of text distance in hierarchical clustering was improved. The experimental results show that the overall similarity of topics in the platform is good, but the focus of topics makes the content of many journals not identifiable, which affects the user's positioning of topics.

收稿日期: 2018−05−30
第一作者: 杨传春(1975−)，男，硕士研究生．研究方向：机器学习．E-mail：47105036@qq.com
通信作者: 郭　强(1975−)，女，教授．研究方向：网络科学、商务智能、科学知识图谱分析．E-mail：qiang.guo@usst.edu.cn

Keywords: *LDA model*; *generating model*; *topic discovery*; *hierarchical clustering*; *text mining*

快速发现和获取网络中的目标信息是自然语言处理、机器学习、人工智能等领域内的研究热点。文本的主题是目标信息的重要载体，获取主题信息是实现准确定位的重要途径。不同的网络数据和处理要求，其发现算法一般有差异[1-3]，但大多是基于LDA(latent Dirichlet allocation)生成模型实现的一种统计方法，该方法可以挖掘文本内词与词之间的语义联系[4]。在此基础上进行的文本聚类能有效地对文本内潜在的主题进行提炼和划分，从而实现缩小查找范围、快速准确定位目标信息的目的。

生成模型的显著优点是，不需要预先对文本进行归类或标记，具有无监督的机器学习特点。文献[5]对具有规范语言特点和明确关键词的社科文献进行了主题建模，提出了主题引导词库的概念；文献[6]把文本聚类应用于查新系统，进行了主题挖掘实验，并对实验效果进行了自评；文献[7]则对新浪平台的微博网络社区进行了主题发现，以微博的主题分布作为衡量微博主题在用户中影响力大小的依据。

近年来，在线网络教育内容呈现出爆炸式增长，其巨量数据的生成与维护成为在线教育吸引大众的重要资源。本文基于LDA文本生成模型，对一个开放的在线教育机构网络刊物进行主题发现与聚类研究，提出了主题发现和聚类的流程，对层次聚类文本距离的算法作出了改进，提出了合并向量算法(union vector method，UVM)。

1 LDA主题模型

1.1 模型概述

LDA生成模型简称LDA模型，是挖掘文本集内文本主题潜在关系的算法，它模拟文本的生成过程。文本的主题发现与文本的生成是个互逆过程，文本的主题发现是从文本中发现词项描述的主题，以及不同主题间的关联，同一文本可以有多个主题；文本的生成是根据主题，从词库内选用与主题相匹配的词项来构成文本。两者的路径相反，通过先验参数，LDA模型实现了文本生成逆问题的解决方案。

假定要生成包含M篇文本的文本集，文本集的内容与K个主题相关，每个文本所用的词项均来自于同一个包含V个元素的集合。文本生成过程如下[8]：

步骤1 生成文本的主题分布$\boldsymbol{\Theta}$，分布的维度为$M\times K$，如图1所示。

	主题1	主题2	…	主题K
文本1	p_{11}	p_{12}	…	p_{1K}
⋮	⋮	⋮	p_{mk}	⋮
文本M	p_{M1}	p_{M2}	…	p_{MK}

图1 文本−主题分布

Fig.1 Doc-topics distribution

图1中的通项p_{mk}($m\leqslant M$，$k\leqslant K$)为第m篇文本选择主题k的概率，每行的概率和为1。文本−主题分布服从狄利克雷分布Dir($\boldsymbol{\Theta}|\boldsymbol{\alpha}$)，$\boldsymbol{\alpha}$为先验参数，又称超参数。文本的主题分布服从多项式分布Mult($\boldsymbol{\vartheta}$)，依据该分布选择概率值最大的主题。

步骤2 生成主题−词项分布$\boldsymbol{\Phi}$，分布的维度为$K\times N_m$，N_m为第m篇文本总的词数量，如图2所示。

	词项1	词项2	…	词项N_m
主题1	p_{11}	p_{12}	…	p_{1N_m}
⋮	⋮	⋮	p_{kn}	⋮
主题M	p_{K1}	p_{K2}	…	p_{KN_m}

图2 主题−词项分布

Fig.2 Topic-words distribution

图2中的通项p_{kn}($k\leqslant K$，$n\leqslant N_m$)为第k个主题下选中第n个词项的概率，每列的概率和为1。主题−词项分布服从狄利克雷分布Dir($\boldsymbol{\Phi}|\boldsymbol{\beta}$)，$\boldsymbol{\beta}$为先验参数。主题下的词分布服从多项式分布Mult($\boldsymbol{\varphi}$)，依据该分布选择概率值最大的词项w_j。

步骤3 迭代上述过程，直至产生文本集。

1.2 模型的数学原理

词袋(bag of words，BOW)中有V个独立同分布的不重复词，取N个词生成一篇文本(文本可取重复的词)，由每个生成后的词在文本中出现的概率，反过来便可得到文本W的生成概率为

$$p(W)=[p(w_1)]^{n_1}[p(w_2)]^{n_2}\cdots[p(w_v)]^{n_v}=\prod_{j=1}^{v}p_j^{(n_j)}\quad(1)$$

式中：$p(w_j)=p_j$；$\sum_{j=1}^{v}p_j=1$；$N=n_1+n_2+\cdots+n_v$，$v\leqslant V$。实验中，文本内每个词的概率可以用出现的次数与该文本总的词项数量之比来估计，即 $p_j=\hat{p}_j=\frac{n_j}{N}(j\in[1,v])$，符号“$\hat{}$”表示估计值。

文本的生成过程可以用图 3 表述[8]。

图 3 中的矩形表示框内相应变量 n，m，k 的作用域；阴影圆圈内的值表示实验中是可以被观测到的已知量，矩形框内无阴影圆圈表示隐含变量；矩形框外的值表示先验参数，即狄利克雷分布的超参数，需在实验前预先给定；$\boldsymbol{\varphi}_k$ 的箭头穿透了 2 层矩形框指向 $w_{m,n}$，表示生成该词项前有一个文本到主题和先验参数 $\boldsymbol{\beta}$ 到主题的混合过程。图中各量及其他相关符号含义见表 1。

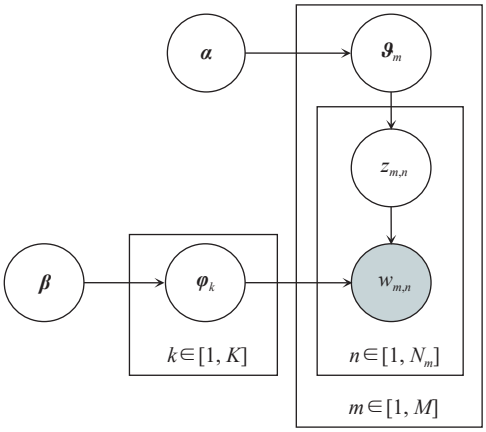

图 3　LDA 混合生成模型

Fig. 3　Model of LDA hybrid generation

文本集中第 m 篇文本的生成过程可分为两步[9-10]：a. 由先验参数 $\boldsymbol{\alpha}$ 获得文本的主题分布 $\boldsymbol{\vartheta}_m=p(z|W_m)$，选取主题；b. 由先验参数 $\boldsymbol{\beta}$ 获得每个主题的词项分布 $\boldsymbol{\varphi}_k=p(t|z=k)$，选取词项。因此，LDA 模型文本生成过程可以看成 3 层条件概率过程。

表 1　使用符号及其含义

Tab.1　Symbles and meanings used

符号	含义
M	文本集中的文本数量，标量。实验中由已知数据给出，$m\in[1,M]$。
K	文本集中的主题数量，标量。实验中自由确定，正整数，$k\in[1,K]$。
V	文本集中不重复的词项总数，标量。
N_m	第 m 篇文本包含的词项数量，不去重，标量。文本集中不同文本该值一般不同。
$\boldsymbol{\Theta}$	文本集的文本−主题分布，向量，维度为 $M\times K$，待估计参数 $\boldsymbol{\Theta}=\{\boldsymbol{\vartheta}_m\}_{m=1}^{M}$。
$\boldsymbol{\Phi}$	文本集的主题−词项分布，向量，维度为 $K\times N_m$，待估计参数 $\boldsymbol{\Phi}=\{\boldsymbol{\varphi}_k\}_{k=1}^{K}$。
$\boldsymbol{\alpha}$	文本−主题分布的先验参数。本实验中各分量均取 $\alpha=60/K$。
$\boldsymbol{\beta}$	主题−词项分布的先验参数。本实验中各分量均取 $\beta=0.01$。
$\boldsymbol{\vartheta}_m$	第 m 篇文本的文本−主题分布，$\sim\mathrm{Dir}(\boldsymbol{\alpha})$，向量，维度为 $1\times K$。
$\boldsymbol{\varphi}_k$	第 k 个主题的主题−词项分布，$\sim\mathrm{Dir}(\boldsymbol{\beta})$，向量，维度为 $1\times N_m$。
$z_{m,n}$	文本集中第 m 篇文本的主题−词项分布，标量。
$w_{m,n}$	第 m 篇文本的第 n 个词，标量。t 取自于词库的词，w 表示文本中的任意一个词。
$\mathcal{W}$。	文本集，$\mathcal{W}=\{W_m\}_{m=1}^{M}$

第一层：基于主题概率的生成过程。第 m 篇文本中词项 t 的生成概率可表示为

$$p(w_{m,n}=t)=\sum_{k=1}^{K}p(w_{m,n}=t|z=k)\,p(z=k)\qquad(2)$$

式中：$\sum_{k=1}^{K}p(z=k)=1$；每个词项 t 服从多项式分布。

第二层：混合层，基于主题概率和基于文本概率的生成过程。第 m 篇文本的生成概率为文本−主题概率与主题−词项概率的乘积

$$p(w_{m,n}=t|\boldsymbol{\vartheta}_m,\boldsymbol{\Phi})=\sum_{k=1}^{K}p(w_{m,n}=t|\boldsymbol{\varphi}_k)\,p(z_{m,n}=k|\boldsymbol{\vartheta}_m)\qquad(3)$$

式中，$\boldsymbol{\varphi}_k$ 的取值由 $p(z_{m,n}=k|\boldsymbol{\vartheta}_m)$ 中的主题值 k 确定。

第三层：基于先验参数的生成过程。文本集 $\mathcal{W}$ 的生成概率等于每篇文本的概率积

$$p(\boldsymbol{\mathcal{W}}|\boldsymbol{\alpha},\boldsymbol{\beta})=\prod_{m=1}^{M}p(W_m|\boldsymbol{\alpha},\boldsymbol{\beta})=\prod_{m=1}^{M}\iint\left(p(\boldsymbol{\vartheta}_m|\boldsymbol{\alpha})\cdot p(\boldsymbol{\phi}|\boldsymbol{\beta})\prod_{n=1}^{N_m}p(w_{m,n}|\boldsymbol{\vartheta}_m,\boldsymbol{\phi})\right)\mathrm{d}\boldsymbol{\phi}\mathrm{d}\boldsymbol{\vartheta}_m=$$
$$\prod_{m=1}^{M}\left(\iint\left(\overbrace{p(\boldsymbol{\vartheta}_m|\boldsymbol{\alpha})}^{\sim\mathrm{Dir}(\boldsymbol{\alpha})}\overbrace{p(\boldsymbol{\phi}|\boldsymbol{\beta})}^{\sim\mathrm{Dir}(\boldsymbol{\beta})}\prod_{n=1}^{N_m}\left(\sum_{k=1}^{K}\overbrace{p(w_{m,n}=t|\boldsymbol{\varphi}_k)}^{\sim\mathrm{Mult}(N_m)}\overbrace{p(z_{m,n}=k|\boldsymbol{\vartheta}_m)}^{\sim\mathrm{Mult}(K)}\right)\right)\mathrm{d}\boldsymbol{\phi}\mathrm{d}\boldsymbol{\vartheta}_m\right) \tag{4}$$

1.3 吉布斯采样

LDA模型在概念上很简单，但要实现求解就有很大困难，一般采用近似计算完成，吉布斯采样是完成模型计算的一个理想方法，它可以消除先验参数值对结果的影响。吉布斯采样是对马尔科夫链蒙特卡洛法（Markov-chain Monte Carlo，MCMC）的一种模拟，通过马尔可夫链的平稳态来模拟高维分布$p(\boldsymbol{x}_i)$，某时刻$\boldsymbol{x}_i$的值只和生成$\boldsymbol{x}_i$之前的状态$\boldsymbol{x}_{\neg i}=(x_1,x_2,\ldots,x_{i-1})$有关（$\neg i$表示不包含$i$状态），即

$$p(z_i|\boldsymbol{z}_{\neg i},\boldsymbol{x})=\frac{p(\boldsymbol{z},\boldsymbol{x})}{p(\boldsymbol{x}_{\neg i},\boldsymbol{x})} \tag{5}$$

式中：$\boldsymbol{x}=\{x_i,\boldsymbol{x}_{\neg i}\}$；$\boldsymbol{z}$为隐含变量。

LDA模型中，文本−主题分布$\boldsymbol{\Theta}$和主题−词项分布$\boldsymbol{\Phi}$均是隐含的变量。它们包含在由两个先验超参数确定的联合分布$p(\boldsymbol{w},\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\beta})$中，文本主题分布和主题词项分布是两个独立过程，因此联合分布可分解成两个独立项

$$p(\boldsymbol{w},\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\beta})=p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta})\,p(\boldsymbol{z}|\boldsymbol{\alpha}) \tag{6}$$

其中$p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta})$由下式给出：

$$p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta})=\int\underbrace{p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\phi})}_{\sim\mathrm{Mult}\left(n_z^{(t)}\right)}\underbrace{p(\boldsymbol{\phi}|\boldsymbol{\beta})}_{\sim\mathrm{Dir}(\boldsymbol{\beta})}\mathrm{d}\boldsymbol{\phi}=$$
$$\int\prod_{z=1}^{K}\frac{1}{Be(\boldsymbol{\beta})}\prod_{t=1}^{V}\varphi_{z,t}^{n_z^{(t)}+\beta_t-1}\mathrm{d}\boldsymbol{\varphi}_z=\prod_{z=1}^{K}\frac{\Delta(\boldsymbol{n}_z+\boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})} \tag{7}$$

式中：$n_z^{(t)}$表示主题−词项分布中主题z出现词项t的次数；$\boldsymbol{n}_z=\left\{n_z^{(t)}\right\}_{t=1}^{V}$；$\Delta(\boldsymbol{\beta})=\dfrac{\prod_{i=1}^{K}\Gamma(\beta_i)}{\Gamma\left(\sum_{i=1}^{K}\beta_i\right)}$是贝塔分布，在整数范围内有$\Gamma(n+1)=n\Gamma(n)$。式（6）中$p(\boldsymbol{z}|\boldsymbol{\alpha})$由下式确定：

$$p(\boldsymbol{z}|\boldsymbol{\alpha})=\int\underbrace{p(\boldsymbol{z}|\boldsymbol{\Theta})}_{\sim\mathrm{Mult}(n_m^{(k)})}\underbrace{p(\boldsymbol{\Theta}|\boldsymbol{\alpha})}_{\sim\mathrm{Dir}(\boldsymbol{\alpha})}\mathrm{d}\boldsymbol{\Theta}=$$
$$\int\prod_{m=1}^{M}\frac{1}{Be(\boldsymbol{\alpha})}\prod_{k=1}^{K}\boldsymbol{\vartheta}_{m,k}^{n_m^{(k)}+\alpha_k-1}\mathrm{d}\boldsymbol{\vartheta}_m=\prod_{m=1}^{M}\frac{\Delta(\boldsymbol{n}_m+\boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})} \tag{8}$$

式中，$n_m^{(k)}$表示文本−主题分布中第m篇文本出现主题k的次数，$\boldsymbol{n}_m=\left\{n_m^{(k)}\right\}_{k=1}^{K}$。因此，式（6）的联合分布$p(\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta})$变为

$$p(\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta})=\prod_{k=1}^{K}\frac{\Delta(\boldsymbol{n}_k+\boldsymbol{\beta})}{\Delta(\boldsymbol{\beta})}\prod_{m=1}^{M}\frac{\Delta(\boldsymbol{n}_m+\boldsymbol{\alpha})}{\Delta(\boldsymbol{\alpha})} \tag{9}$$

根据式（5）可得主题吉布斯采样更新规则[11]：

$$p(z_i=k|\boldsymbol{z}_{\neg i},\boldsymbol{w})=\frac{p(\boldsymbol{w},\boldsymbol{z})}{p(\boldsymbol{w},\boldsymbol{z}_{\neg i})}=$$
$$\frac{p(\boldsymbol{w}|\boldsymbol{z})}{p(\boldsymbol{w}_{\neg i}|\boldsymbol{z}_{\neg i})\,p(w_i)}\frac{p(\boldsymbol{z})}{p(\boldsymbol{z}_{\neg i})}\propto$$
$$\frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{i=1}^{V}n_{k,\neg i}^{(t)}+\beta_t}\frac{n_{m,\neg i}^{(t)}+\alpha_k}{\left[\sum_{k=1}^{K}n_m^{(k)}+\alpha_k\right]-1} \tag{10}$$

最后可获得多项式参数$\boldsymbol{\Theta}$和$\boldsymbol{\phi}$的表示，结合多项式分布为狄利克雷的后验分布可得

$$p(\boldsymbol{\vartheta}_m|\boldsymbol{\alpha},\boldsymbol{z},\boldsymbol{w})=\frac{1}{Z_{\boldsymbol{\vartheta}_m}}\prod_{n=1}^{N_m}p(z_{m,n}|\boldsymbol{\vartheta}_m)\,p(\boldsymbol{\vartheta}_m|\boldsymbol{\alpha})=$$
$$\mathrm{Dir}(\boldsymbol{\vartheta}_m|\boldsymbol{n}_m+\boldsymbol{\alpha}) \tag{11}$$

$$p(\boldsymbol{\varphi}_k|\boldsymbol{\beta},\boldsymbol{z},\boldsymbol{w})=\frac{1}{Z_{\boldsymbol{\varphi}_k}}\prod_{\{i:z_i=k\}}^{N_m}p(w_i|\boldsymbol{\varphi}_k)\,p(\boldsymbol{\varphi}_k|\boldsymbol{\beta})=$$
$$\mathrm{Dir}(\boldsymbol{\varphi}_k|\boldsymbol{n}_k+\boldsymbol{\beta}) \tag{12}$$

根据狄利克雷分布的期望$\langle\mathrm{Dir}(\boldsymbol{\alpha})\rangle=a_i/\sum_i a_i$，应用于式（11）和式（12）可得[11]

$$\varphi_{k,t}=\frac{n_k^{(t)}+\beta_t}{\sum_{i=1}^{V}n_k^{(t)}+\beta_t},\vartheta_{m,k}=\frac{n_m^{(k)}+\alpha_k}{\sum_{k=1}^{K}n_m^{(k)}+\alpha_k} \tag{13}$$

式（9）是LDA模型数学原理的核心，它通过式（7）和式（8）包含了文本−主题分布和主题−词项分布，文本主题发现的信息就集中在这两个概率分布中。然而由该式计算出来的结果具有不确定性，因为式中包含有先验参数成份，因此需要不断迭代以获得稳定的文本−主题分布和主题−词项分布，这两个分布更新规则通过吉布斯采样求得的式（13）确定，这样LDA生成模型就确定下来了。

1.4 困惑度

评价LDA生成模型质量的标准一般用困惑度表示，困惑度的定义为[8]

$$Perplexity = P(\mathcal{W}|\boldsymbol{w},\boldsymbol{z}) = \prod_{m=1}^{M}(p(\boldsymbol{w}_m|\boldsymbol{w},\boldsymbol{z}))^{-\frac{1}{N}} = \exp\left(-\frac{\sum_{m=1}^{M}\ln(p(\boldsymbol{w}_m|\boldsymbol{w},\boldsymbol{z}))}{\sum_{m=1}^{M}N_m}\right) \tag{14}$$

式中：$\mathcal{W}=\{W_m\}_{m=1}^{M}$；$\boldsymbol{w}_m=\{w_{m,n}\}_{n=1}^{N_m}$、$\boldsymbol{z}=\{z_k\}_{k=1}^{K}$；$N=\sum_{m=1}^{M}N_m$，且有

$$p(\boldsymbol{w}_m|\boldsymbol{w},\boldsymbol{z})=\prod_{n=1}^{N_m}\sum_{k=1}^{K}(p(w_{m,n}=t|z_n=k)\,p(z_n=k|W=m))=\prod_{t=1}^{V}\left(\sum_{k=1}^{K}\varphi_{k,t}\vartheta_{m,k}\right)^{n_m^{(t)}}$$

由此可得，

$$Perplexity=\exp\left(-\frac{\sum_{m=1}^{M}n_m^{(t)}\left(\sum_{t=1}^{V}\ln\left(\left(\sum_{k=1}^{K}\varphi_{k,t}\vartheta_{m,k}\right)\right)\right)}{\sum_{m=1}^{M}N_m}\right)$$

式中，$n_m^{(t)}$表示词项 t 在文本 m 中出现的次数。从定义式$\prod_{m=1}^{M}(p(\boldsymbol{w}_m|\boldsymbol{w},\boldsymbol{z}))^{-\frac{1}{N}}$可以看出，困惑度越小，表示模型生成文本集的概率越大，模型的生成效果越好。

2　聚　类

2.1　TF-IDF

文本的聚类就是挖掘文本的相似性，是以描述文本的主题信息为基础的，信息重要性的度量常用 TF-IDF 算法[12]。不同的词在文本中出现的次数一般不同，可以用词项在文本中出现的次数词频（term frequency，TF）表示。文本集中文本数量除以包含某个词的文本数量叫逆文本频率（inverse document frequency，IDF），该值越大，表示词项在文本集内具有一定的“独特性”。

$$TF_n=\frac{N_{m,n}}{N_m} \tag{15}$$

式中：分子 $N_{m,n}$ 表示第 m 篇文本中第 n 个词在该文本中出现的次数；分母 N_m 表示第 m 篇文本中所有词项的数量和。逆文本频率为

$$IDF_n=\log\frac{M}{\{m:w_n\in W_m\}} \tag{16}$$

式中：M 为文本集$\mathcal{W}$中文本的总数量；分母表示包含第 n 个词项 w_n 的文本数量。两者的比值取对数是为了平衡 TF 和 IDF 对同一词项 w_n 的影响。

2.2　文本距离

文本距离用来反映文本的相似程度，距离越小表示文本越相似。要实现文本距离的计算，先要实现文本的计算表示，向量空间模型（vector space model，VSM）就是把文本语义的相似度简化为空间向量运算的模型[13]，它以词袋为基础，把文本的每个关键与词袋对比，赋予一个正实数值，使每篇文本构成一个多维空间向量，从而实现数据的计算表示。文本间距离的度量常用的有余弦相似度和杰森–香农距离（Jensen-Shannon distance，JSD）[14]，两者都具有对称性。余弦相似度的定义为

$$\cos\theta=\frac{\sum_{i=1}^{n}(A_iB_i)}{\sqrt{\sum_{i=1}^{n}A_i^2}\sqrt{\sum_{i=1}^{n}B_i^2}} \tag{17}$$

余弦相似度为零，表示两个向量无语义关联性；相似度越接近于 1，表示文本越相似。JSD 的定义为

$$D_{\mathrm{JS}}(\boldsymbol{A},\boldsymbol{B})=\frac{1}{2}D_{\mathrm{KL}}\left(\boldsymbol{A},\frac{\boldsymbol{A}+\boldsymbol{B}}{2}\right)+\frac{1}{2}D_{\mathrm{KL}}\left(\boldsymbol{B},\frac{\boldsymbol{A}+\boldsymbol{B}}{2}\right) \tag{18}$$

式中：$D_{\mathrm{KL}}(\boldsymbol{A},\boldsymbol{M})=\sum_{x\in X}\left(\boldsymbol{A}(x)\ln\frac{\boldsymbol{A}(x)}{\boldsymbol{M}(x)}\right)$；$\boldsymbol{A}$，$\boldsymbol{B}$ 为基于词袋模型的 VSM 向量值；$\boldsymbol{M}$ 是与 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 具有相同维度的向量。

3　实验过程与结果分析

本实验的主题发现和聚类流程如图 4 所示，实线箭头所指的方向为实验时间顺序和步骤，虚线所指表示需要的相关算法或技术路线。

3.1　数据抓取与清洗

通过分析在线教育机构的网页特点，利用 python 语言的 urllib，urllib2 模块和正则表达式制作爬虫程序，匹配目标字段抓取数据，共取得了 2 794 篇网络刊物的名称、分类和概述 3 个字段值。刊物内容涉及学习和兴趣的各个方面。

由于页面中存在着大量不规范的词语，抓取下来的数据无法被直接利用，必须进行词语替

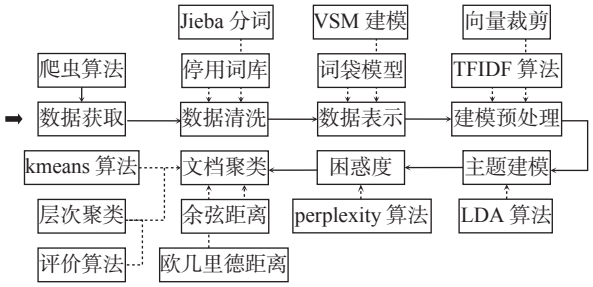

图 4 文本聚类流程和技术路线

Fig. 4 Text clustering process and technical route

换、去污和深度清洗，仅保留简体中文数据。通过建立非目标字符库，构造字符过滤器，把文本集中的数据与过滤器字符库逐一对比达到清洗的目的。采用开源的结巴(Jieba)分词算法，对文本集的全部语句进行分词便可得到需要的数据。

3.2 数据表示与建模预处理

清洗后的词项构成了实验的文本集，通过去重建立词袋，将词袋与每篇文本所含词项对比，构造包含 0 与 1 的多维向量空间，1 表示文本与词袋有映射关系，0 表示文本内没有该词。每个文本对应一个向量，向量的分量按词袋内词的顺序排列，形成 1110010101…10 结构序列，序列长度等于文本内词项总数。计算各词的 TF-IDF 值，用其置换向量序列内的 1，实现向量的计算表示。

计算文本间的余弦相似度，如图 5 所示，它反映了不同余弦相似度随文本数量变化的关系。随着余弦值的增加，具有相似性的文本数量逐渐减少，没有完全相同的文本(*cosine*=1)，其中的子图是母图文本数量归一化后的情况。

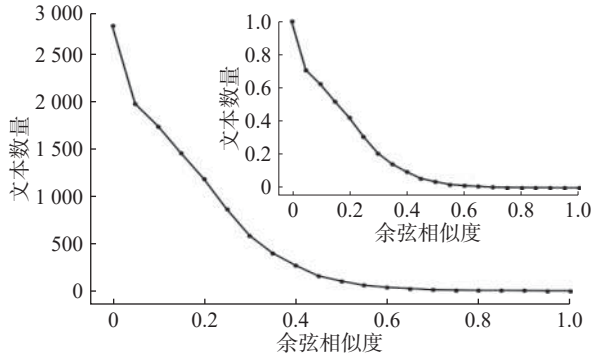

图 5 余弦相似度与文本数量关系

Fig.5 Relation of cosine similarity and text quantity

3.3 主题建模

LDA 模型的输入值有，自定义主题数量 K(60～440)、文本主题分布的超参数 α=0.01、主题词项分布超参数 β=60/K，模型迭代的次数为 1 000 次；模型的输出有文本主题分布 $\boldsymbol{\Theta}$、主题词项分布 $\boldsymbol{\phi}$，以及词袋与自定义特征整数间的映射表，该映射表并非必须，但通过它易于人机交互计算建模。

利用先验参数和文本集对模型作持续训练，同时以 20 为步长改变主题数，得到困惑度变化如图 6 所示。数据显示，当主题数为 320 时，模型困惑度最小，表明模型此时的生成效果最好；K>320 时，模型的困惑度基本保持不变，这有可能是模型陷入了局部极小值的原因。困惑度最小意味着主题生成的概率最大。

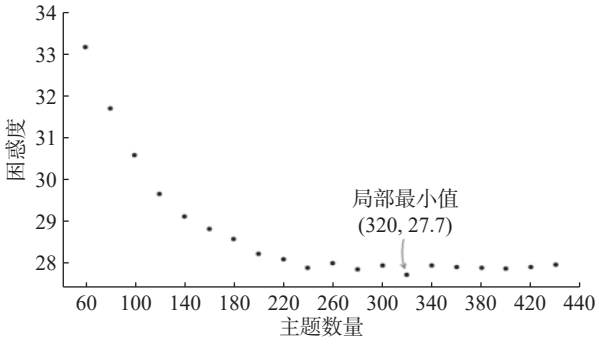

图 6 困惑度

Fig.6 Perplexity

表 2 显示了主题数 K=320 时，模型提取出的前 4 个主题和相应的生成概率。可以看出每个主题的关键词聚焦的内容语义边界非常明显，表明 LDA 模型挖掘出了文本主题间的内在联系。与之对比，当主题数目 K=80 时，模型提取的前 4 个主题和相应的生成概率如表 3 所示。此时，主题的关键词之间有些模糊，主题边界出现了交叉现象，交叉词语已用下划线标出。这种情况有可能是模型在进行数据采样时，未进入平稳态程序就已经结束。实验表明，当迭代次数增加到 2 000 次时，模型输出的主题结果并未改变，说明即使迭代只有 1 000 次时，模型事实上已进入了采样的平稳态。因此只能表明，如果把全部在线网络刊物的内容归纳到现有的主题数目时，主题定位将不再清晰，需要进一步细分或者放大主题内涵才能确定更清晰的主题边界。

3.4 层次聚类及评价

聚类时，将每个刊物作为一个独立的簇，计算两两文本之间的距离，并设定距离阈值簇的容量，依次合并不同的簇，合并后的簇以簇内所有样本点的距离平均值作为迭代计算的依据，反复迭代直到最后所有簇均被合并[15]。这一算法只有在能定义簇平均值时才有意义，其时间复杂度为 $O(nmkt)$，其中 n 为数据点的数目，m 为数据点的维度，k 为簇的数目，t 为迭代次数[16]。

本实验中，合并以后的簇采用簇内各数据点向量的并集来表示簇，并以此计算簇间的距离，该算法称之为合并向量算法（UVM），算法的伪码如表 4 所示。

UVM 算法与文献[15]算法随词袋中词项数量的性能对比如图 7 所示。

聚类结果如图 8 所示，横轴为各个刊物的序号，纵轴为向量间的欧几里德距离。最后 15 步的聚类如图 9 所示，横轴上带括号的数字表示该聚类叶子包含的刊物数量，未加括号的表示刊物序号。从图 9 可以简单统计出，最后一步的聚类中，两个分支的刊物数量分别为 4 和 2 790，表明

表 2　主题和词项的概率（*K*=320）

Tab.2　Probability of topics and items（*K*=320）

主题 1	概率	主题 2	概率	主题 3	概率	主题 4	概率
学习	0.208 777	电影	0.435 958	日语	0.438 138	世界	0.439 402
新概念	0.093 978	人生	0.078 881	零基础	0.221 856	国家	0.048 275
模式	0.052 233	电视剧	0.037 145	交流	0.124 806	角落	0.021 485
时间	0.047 015	影评	0.027 871	学习	0.041 620	圆心	0.021 485
改变	0.047 015	影像	0.023 233	沪江日语	0.036 075	母语	0.016 127
思维	0.047 015	电视	0.023 233	资料集	0.005 573	地理	0.016 127
沪江	0.041 797	飞鱼	0.023 233	造句	0.005 573	超越	0.016 127
习惯	0.036 579	电影学	0.018 596	菊花	0.005 573	谣言	0.016 127
信心	0.026 143	解说	0.018 596	动漫	0.002 801	漫游	0.010 769
动力	0.020 925	盘点	0.018 596	游戏卡	0.002 801	絮语	0.010 769

表 3　主题和词项的概率（*K*=80）

Tab.3　Probability of topics and items（*K*=80）

主题 1	概率	主题 2	概率	主题 3	概率	主题 4	概率
生活	0.278 708	语种	0.535 788	**学习**	0.729 851	**职场**	0.464 959
兴趣	0.226 169	**学习**	0.277 830	订阅	0.051 344	**兴趣**	0.366 958
心理	0.207 895	英语听力	0.059 557	游戏	0.039 941	精选	0.025 093
中国	0.070 836	发现	0.022 706	出品	0.017 134	原创	0.013 697
职场	0.061 699	英国	0.011 367	关注	0.014 283	国家	0.013 697
时间	0.045 709	学霸	0.008 532	美文	0.014 283	精品	0.011 418
科学	0.009 160	日剧	0.005 698	伙伴	0.014 283	交流	0.009 139
真实	0.009 160	笑话	0.005 698	动态	0.011 432	杂志	0.009 139
心理咨询	0.006 876	粘土	0.005 698	粤语	0.011 432	短片	0.006 860
水平	0.004 591	圣诞节	0.005 698	专辑	0.008 581	版本	0.006 860

表 4　UVM 算法伪码表示

Tab.4　Representation of pseudo-code of UVM algorithm

step1: generate the bag of words BOW=$\{w_1,w_2,\ldots,w_V\}$

step2: for doc_i in docset:

create the VSM_i of the ith document

step3: calculating the cosine simularity between every 2 documents pair as input value of clustering

step4: *num* = length of documents set - 1

if *num* >1

clustering C_i、C_j…into a new cluster　according to the *VSM* value

$$VSM_C=VSM_i\cup VSM_j\cup\cdots$$

$$num=num-n_{VSM_i}-n_{VSM_j}-\cdots$$

goto step 3

V represent the number of unique words in document set.

W is a document can be added sub marker such as a letter

w is a word which is unique in BOW,

but it can be repeated in a *W*

num is the cluster number after clustering

C_i etc. means a cluster which can include different critical topics

$$n_{VSM_i}=n_{VSM_j}=\cdots=1$$

全部刊物的主题总体上比较集中，有利于用户在相应主题下通过聚类层次图向下找到更为确切的主题，实现主题定位。但是，平台主题的过度集中，也暴露出刊物数量相对于内容的冗余，如果平台能对主题相似度较大的刊物进行归并，适当缩减刊物数量，精炼刊物主题，不仅能增强不同刊物在用户使用中的辨识度，也可以节约大量的平台人力维护成本。

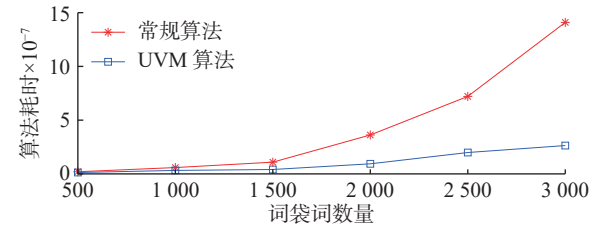

图 7 UVM 算法性能

Fig.7 UVM algorithm performance

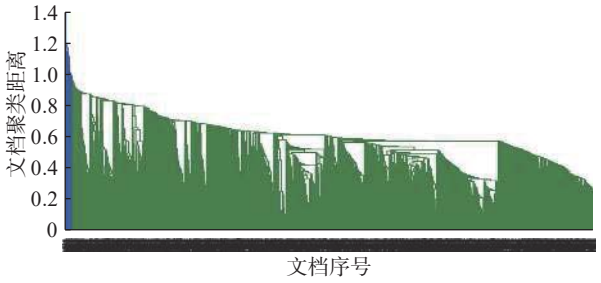

图 8 全部刊物层次聚类图

Fig.8 Cluster graph of total journals

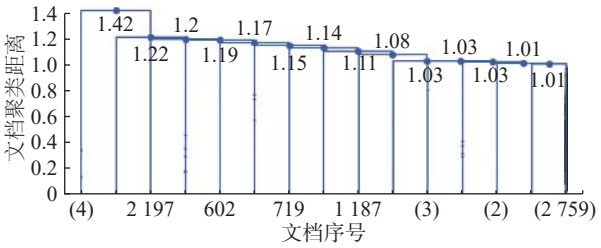

图 9 最后 15 步聚类图

Fig.9 Graph of last 15 steps

4 结束语

对在线网络教育平台的刊物进行了主题发现和聚类研究。实验数据通过爬虫获取，提出了主题发现和聚类的一般流程，通过对数据的清洗、生成词袋、计算 TF-IDF 值和向量空间模型的表示等为主题建模提供了有效数据。以 LDA 生成模型为基础，完成了对 2 794 个刊物的潜在主题的发现，以文本的关键词为基础，利用层次聚类模型进行了主题聚类工作，并提出了新的合并向量算法(UVM 算法)。

实验中 2 794 篇文本中包含 3 800 左右的关键词，聚类难度相当大。实验中采用了分步聚类方法，先用 kmeans 算法聚类成 320 个主题，再进行层次聚类。结果表明，合理的聚类有助于用户快速发现目标信息，在线教育机构亦需要对刊物内容和刊物构成进行适当的优化。

本实验研究的不足之处是文本的层次聚类算法中距离的计算方法，这也是目前聚类算法面临的系统性难题。在计算距离时，无论是采用余弦距离、欧几里德距离还是香农距离等都无法避免极端情况下，距离数值相等或相近引起结果可能出现的大偏差。如表 5 所示，向量 1 与向量 2、向量 2 与向量 3.....向量 n−1 与向量 n 的“距离”是相等的，在进行簇合并时，它们将会被归为同一簇。事实上，向量 1 和向量 n 在语义上可能根本没有相似性，因此，对聚类距离的研究值得继续深入。

表 5 向量在极端情况下的分布

Tab.5 Distribution of vectors in extreme conditions

向量	分布											
向量 1	1	1	1	0	0	0	0	0	0	0	0	0
向量 2	0	1	1	1	0	0	0	0	0	0	0	0
向量 3	0	0	1	1	1	0	0	0	0	0	0	0
⋮						⋮						
向量 n−1	0	0	0	0	0	0	0	0	1	1	1	0
向量 n	0	0	0	0	0	0	0	0	0	1	1	1

参考文献：

[1] 林奕欧, 雷航, 李晓瑜, 等. 自然语言处理中的深度学习: 方法及应用[J]. 电子科技大学学报, 2017, 46(6): 913–919.

[2] 李村合, 朱红波. 基于半监督学习的多示例多标记 E-MIMLSVM+算法 [J]. 计算机工程与应用, 2018, 54(2): 149–154.

[3] 许坤, 冯岩松, 赵东岩, 等. 面向知识库的中文自然语言问句的语义理解[J]. 北京大学学报(自然科学版), 2014, 50(1): 85–92.

[4] 张培晶, 宋蕾. 基于 LDA 的微博文本主题建模方法研究述评[J]. 图书情报工作, 2012, 56(24): 120–126.

[5] 李昌亚, 刘方方. 基于 LDA 的社科文献主题建模方法[J]. 计算机技术与发展, 2018, 28(2): 182–187.

(下转第 306 页)