【信息技术】

LDA模型的优化及其主题数量选择研究

——以科技文献为例

王婷婷 韩 满 王 宇

【摘 要】[目的]为提升传统LDA模型的主题识别性能,并给主题最优数目选择提供技术方案,提出基于自适应聚类的K-wrLDA模型。[方法]利用LDA和Word2Vec模型得出包含主题词概率信息及词义相关性的T-WV矩阵,并将传统LDA模型的主题数目选择问题转化为聚类效果评价问题,以内部指标伪F统计量作为目标函数,计算主题聚类数目的最优解,并对新旧两种模型的主题识别效果进行比较。[结果]经自适应聚类得出最优主题数量为33,且新模型的困惑度得分始终低于传统模型,主题识别效果对比显示新模型具有更好的凝聚性。[局限]在实证语料选取上获取单一主题下的科技文献,数据量不大。[结论]新模型具有更理想的主题识别能力,并能够自主计算最优主题数目。该模型作为对传统LDA模型的改进,可以应用于各领域的大规模语料中。

【关 键 词】主题模型;词嵌入;自适应聚类;困惑度

【作者简介】王婷婷(通讯作者)(ORCID:0000-0001-8008-4627),华侨大学统计学院,华侨大学现代应用统计与大数据研究中心,E-mail:wantting62@126.com;韩满,华侨大学统计学院,华侨大学现代应用统计与大数据研究中心;王宇,华侨大学统计学院(厦门 361021)。

【原文出处】《数据分析与知识发现》(京),2018.1.29~40

【基金项目】本文系国家社会科学基金项目“基于LDA模型的‘海上丝绸之路’文本挖掘研究”(项目编号:15CTJ005)的研究成果之一。

1 引言

LDA主题模型是数据挖掘尤其是文本挖掘和信息处理方面不可或缺的文本建模模型。该模型在具有可靠的数学基础的同时便于拓展应用,因此自提出以来就受到广大学者的青睐。截至目前,有关LDA模型的原始文献的引用数量已经达到17 215次(数据来源于Google学术搜索,检索时间2017年1月22日)。其应用范围包括:学术文献挖掘、社交媒体短文本分析、情感倾向性分析、自然语言处理、网络结构数据挖掘等诸多方面。

LDA模型的广泛应用具有其深刻的内部原因。首先,在主题模型问世之前,有关文本的表示方法会将文本转化成高维稀疏矩阵,例如应用最多的向量空间模型(VSM)就是这种处理手段,将非结构化的文本数据转化为超高维的结构化数据。但这种处理方法容易引起“维数灾难”,给后续的计算分析带来一定困扰。其次,随着人们对文本数据的重视,对海量文本深层含义的理解提出更高要求。面对纷繁复杂的海量文本数据,如何在相对短的时间内掌握文本内涵,是大众的迫切需求。LDA模型打破了传统文本表示的思维定式,提出“主题”的概念,用于表示文档的信息浓缩,在维度压缩的同时使数据的表现力增强。因此,LDA模型在文本挖掘分析中一直保持着较高的热度。

2 文献综述

Blei等[1]首次提出LDA(Latent Dirichlet Allocation)主题模型,这是一种三层贝叶斯结构,该模型的出现完成了主题模型在贝叶斯层面的拓展并取得广泛的应用。在应用的同时,原作者以及许多学者对LDA模型进行了各种改进与拓展,并将其应用在不同领

域。主题相关性方面，提出传统LDA模型的原作者将参数分布由Dirichlet改为Logistic，给出相关主题模型(Correlated Topic Model，CTM)[2]，以解决传统模型的词袋问题。由此可见，修改参数分布是一种解决思路，而本文采用词嵌入形式解决主题相关性这一问题，则是另一种可行途径。此外，在之前多种多样的理论与实证分析研究中，LDA模型的有效性和可靠性得到充分证明，但LDA主题模型中主题个数的选择问题，依然没有得到有效解决。主题个数的选择直接影响到LDA模型对文本数据的释义情况和主题识别效果，因此非常有必要对这一问题加以解决。由于该问题非常重要，因此国内外学者均有涉猎，主要方法有以下几种：

(1)具有启发式的经验设定法。主题个数K的选择相当于模型评估问题，而对于模型的评估非常困难，因此部分研究者采用具有启发式的经验设定法。他们通过反复调试进行经验性的主观判断，从而确定主题个数。该方法简单、操作性强，在现实中最为常用。关鹏等[3]对科学文献语料库进行LDA模型的主题抽取效果评价过程中，主题个数的确定就采用上述方法。

(2)贝叶斯统计标准方法。该方法首先由Griffiths等[4]提出，随即成为一种确定主题数量的方法。石晶等[5]以及Hajjem等[6]分别基于LDA模型在文本分割和微博信息过滤方面展开应用，均采用贝叶斯统计标准方法进行主题数目K的确定。但这种方法依然是半启发式的，借助经验值与Gibbs抽样算法完成，计算复杂度高。

(3)困惑度(Perplexity)指标。Blei等[1]提出困惑度这一概念，并将其作为模型评判指标。部分学者通过最小化困惑度指标选取主题数目，例如廖列法等[7]和刘江华[8]。但困惑度指标反映模型本身的泛化能力，仅能说明模型对新样本的适用性，以此判定主题数缺乏逻辑严谨性。

(4)非参数方法。这种方法的主要思想就是对主题个数进行非参数化的变形，从而达到在模型运算过程中，无须人为干预，自主学习出最优主题数目。其中比较有代表性的就是Teh等[9]的基于狄利克雷过程所提出的HDP(层次狄利克雷)方法。颜端武等[10]和唐浩浩等[11]即是采用HDP法进行主题数量确定。该方法排除了启发式方法的主观性问题，但会使其他超参数的设定变得更加复杂。另外，此方法计算复杂度高、代码维护成本偏大。

(5)其他方法。除了上述方法之外，一些学者从主题之间相似度的刻画出发，通过构建度量主题相似度指标进行主题数目的确定。其中较常用的刻画指标为KL散度、余弦相似度、JS散度等。曹娟等[12]和关鹏等[13]均采用类似的方法进行最优主题数目的确定。但最优主题数目对不同的相似度度量手段以及指标构造方法都相对敏感，因此指标选取和构造方式的主观性会直接影响主题数目的最终选择。

鉴于传统LDA模型遵循词袋假设，词语之间的相关性被忽略，因此本文在传统LDA模型的基础上，提出一种考虑词语相关关系的新型主题模型，并兼顾主题个数的内生化选择。思路如下：利用Word2Vec模型在探索词义相关关系方面的优越性能，进一步明确主题语料之间的隐含语义关系，把LDA模型中传统的T-W(主题-词)主题分布矩阵变成具有相关关系表达的T-WV(主题-词向量)矩阵。并在此基础上引入自适应聚类方法，构建自适应聚类算法的目标函数，使其在给定的参数范围内可获得局部最优解。在此进行迭代的数据矩阵不同于传统的原始语料，是具有词语相关性的量化信息以及主题词概率大小的排序信息，优化后的数据可使整个聚类过程更加有效，并能够将主题个数的选择转化为聚类评价，在自适应聚类过程中直接解决最优主题数目的判别问题。

综上，本文提出基于自适应聚类的K-wrLDA(即：自适应聚类下的词嵌入相关LDA)模型，优点为：保留了原始矩阵主题词的概率信息的基础上，增强主题词表达性的理解能力，提高同义、近义词之间识别程度，以此提升传统LDA模型的主题划分与识别性能；从多元统计分析的聚类视角改进模型，使得新模型本身具有更好的泛化能力，并解决了传统LDA模型主题聚类个数的选择问题。

3 模型与算法

3.1 基本思路与流程

(1)对原始文本语料进行结构化处理、清洗、降维

等工作，这也是任何文本数据处理过程中必不可少的数据预处理环节。

(2)T-WV矩阵的训练工作，其中包含两个子模块，LDA模块和Word2Vec模块。LDA模型可完成对语料进行深度语义挖掘以及在语义维度层面上的压缩工作，得到T-W(主题-词)的概率分布矩阵；Word2Vec模型可从原始语料中获取词语之间的相关性，并得出词语的向量化表示，进一步得到T-WV(主题-词向量)矩阵。

(3)构建自适应聚类的K-wrLDA模型，再通过设定目标函数进行自适应聚类，确定新模型主题的最优个数，并采用困惑度指标及主题识别的实证结果对传统LDA模型与新模型进行比较。

本文的主要思路流程如图1所示。

图1显示了自适应聚类下的K-wrLDA算法框架。综合而言，这一过程的优势在于：一方面，在没有损失T-W矩阵主题词概率排序信息的情况下，强化了主题词之间的相似性关系，得到映射到高维空间的T-WV矩阵；另一方面，通过自适应聚类中目标函数的构建，有效解决了主题个数的选择问题，提高了模型的泛化能力和主题识别能力。

3.2 基于自适应聚类的K-wrLDA模型

该模型的技术路线主要通过以下三个子模块实现。

(1)T-W矩阵获取过程

采用传统LDA模型，该模型起源于隐性语义索引(LSI)模型，并经历了概率潜在语义分析(pLSI)模型的阶段，属于层次贝叶斯模型的一种。它基于词袋假设认为构成文档的词相互独立，与其出现位置无关。LDA模型中的每个主题实质上是词集上的多项式分布，相同词汇在不同主题下具有不同的概率值。其数学形式如公式(1)[4]所示。

$$p(w_{m,n},z_{m,n},\vec{\theta}_m,\Phi|\vec{\alpha},\vec{\beta})=\prod_{n=1}^{N_m}p(w_{m,n}|\vec{\varphi}_{z_{m,n}})\cdot p(z_{m,n}|\vec{\theta}_m)\cdot p(\vec{\theta}_m|\vec{\alpha})\cdot p(\Phi|\vec{\beta}) \quad (1)$$

这是文本d_m的联合分布形式[3]，其中$z_{m,n}$表示d_m的第n个词项对应的主题，$w_{m,n}$表示第m篇文档的第n个词项；$\vec{\alpha}$、$\vec{\beta}$分别表示d_m的主题分布和某一具体主题$z_{m,n}$词项分布的先验分布；$\vec{\theta}_m$为从$\vec{\alpha}$中抽取d_m的主题分布，是一个K维向量；$\vec{\varphi}_{z_{m,n}}$则为从$\vec{\beta}$中抽取主题$z_{m,n}$对应的词项分布，通过$\vec{\varphi}_{z_{m,n}}$词项$w_{m,n}$被最终确定下来，$\varphi=\{\vec{\varphi}_k\}_{k=1}^{K}$为每个主题的词项分布矩阵。由此构造似然函数[4]，如公式(2)所示。

$$\iint p(\vec{\theta}_m|\vec{\alpha})\cdot p(\Phi|\vec{\beta})\cdot\prod_{n=1}^{N_m}p(w_{m,n}|\vec{\varphi}_{z_{m,n}})p(z_{m,n}|\vec{\theta}_m d\Phi d\vec{\theta}_m \quad (2)$$

但极大似然估计方法并不能求解该问题，Blei等给出的是EM-变分算法，Griffiths等随后提出的Collapsed Gibbs Sampling方法使得模型推导和参数求解极为简化，因此受到普遍推崇。本文所采取的参数估计框架也是基于Gibbs采样展开，这是一种应用非常广泛的蒙特卡洛马尔科夫链(MCMC)模拟方法[14]。

估计先验分布参数$\vec{\alpha}$、$\vec{\beta}$，获取主题分布$\vec{\theta}_m$及词项分布$\vec{\varphi}_{z_{m,n}}$，其中：$\vec{\alpha}\rightarrow\vec{\theta}_m\rightarrow\vec{z}_m$表示生成文本中所有词对应的主题，$\vec{\alpha}\rightarrow\vec{\theta}_m$对应Dirichlet分布部分，$\vec{\theta}_m\rightarrow\vec{z}_m$对应多项式分布部分。利用Dirichlet分布期望可得在迭代过程中更新参数的表达式[4]，如公式(3)所示。

$$\theta_{m,k}=\frac{n_m^{(k)}+\alpha_k}{\sum_{k=1}^{K}n_m^{(k)}+\alpha_k},\ \varphi_{k,t}=\frac{n_k^{(t)}+\beta_t}{\sum_{t=1}^{V}n_k^{(t)}+\beta_t} \quad (3)$$

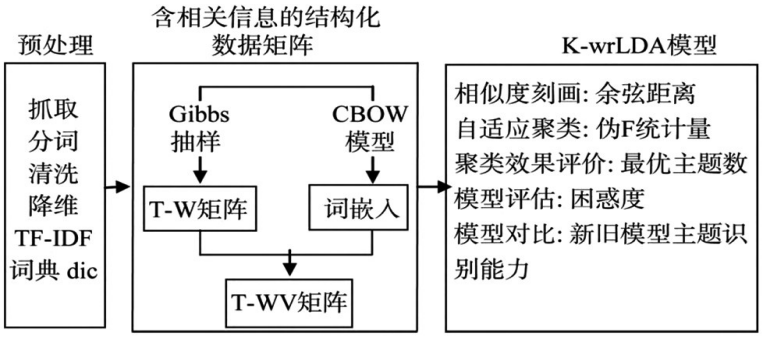


图1 自适应聚类下的K-wrLDA算法框架

其中，$n_m^{(k)}$ 表示第m篇文档 d_m 中第k个主题下词的个数。

在LDA模型中，当采样迭代次数超过一定阈值之后，其参数估计结果可以认为是模型的解，因此Gibbs采样则是对文本每个词项以K维主题为路径采样，并按数值排序的过程，最终主题的条件概率[4]如公式(4)所示。

$$p(z_i=k, w_i=t|\vec{z}_{-i}, \vec{w}_{-i}) \propto \frac{n_{m,-i}^{(k)}+\alpha_k}{\sum_{k=1}^{K} n_{m,-i}^{(k)}+\alpha_k} \cdot \frac{n_{k,-i}^{(t)}+\beta_t}{\sum_{t=1}^{V} n_{k,-i}^{(t)}+\beta_t} \quad (4)$$

以上过程可以得到T-W矩阵①，其算法流程如下：

①D-T(文档-主题)层面：

1)产生主题向量的分布：$\vec{\theta}_m \sim Dir(\vec{\alpha})$；

2)产生文档-主题分布：

$Z_{m,n} \sim Mult(\vec{\theta}_m), Z_{m,n} \in \{1,2,\cdots,K\}$。

②T-W(主题-词)层面：

1)产生词向量的分布：$\vec{\varphi}_k \sim Dir_V(\vec{\beta}_k)$；

2)产生主题-词分布：

$W_{m,n} \sim Mult(\vec{\varphi}_k), W_{m,n} \in \{1,2,\cdots,V\}$。

由于汉语言相近词义较多，单纯地词义挖掘很容易使原本含义相关词的相似程度得不到数学意义的体现。而且LDA模型建立在“词袋模型”的假设基础之上，认为每个词都独立存在，这样的强假设不符合实际，会忽略词语之间的关联性，是LDA模型缺点的根源所在。因此考虑在传统T-W矩阵的基础上，纳入Word2Vec模型，寻找包含词语间相互关系的T-WV矩阵，从而作为弥补LDA模型在词语间相似度刻画方面缺陷的一个有效解决方案。

(2)T-WV矩阵构造过程

对于T-WV矩阵的构造，本文采用Word2Vec模型，这是2013年由Google公司开放的一款将词表征为实数向量的表示学习工具，它所采用的向量表示方式为Hinton[15]提出的Distributed Representation分布式表达。相对LDA模型，Word2Vec模型兼顾了上下文信息，对语义的理解更加准确，因此这种表示方式的优点在于让两个含义相近的词在数学层面具有更高的相似度。Word2Vec模型中最为重要的两个模型分别为CBOW(Continuous Bag-Of-Words)模型和Skip-gram(Continuous Skip-gram)模型。本文主要基于负采样设计CBOW模型，以开展词的向量化工作。CBOW模型是一个三层网络结构，包含输入层、投影层和输出层。

①输入层：包含Context(w)中2c个词的词向量v(Context(w)$_1$)，v(Context(w)$_2$)，…，v(Context(w)$_{2c}$)，其中m为词向量长度。

②投影层：将输入层的2c个向量做求和累加，即 $X_w=\sum_{i=1}^{2c} v(Context(w)_i)$。

③输出层：输出层对应Huffman树，叶子节点共N(=|D|)个，对应词典D中的词，非叶子节点N-1个。

CBOW模型的目标函数[16]如公式(5)所示。

$$L=\sum_{w\in C} \log p(w|Context(w)) \quad (5)$$

其中，$P(w|Context(w))=\frac{e^{y_{w,i_w}}}{\sum_{i=1}^{N} e^{y_{w,i}}}$，$i_w$表示词w在词典D中的索引。

公式(5)若采用梯度下降求解将具有较高的复杂度，通常用Huffman树定义log p(w|Comtext(w)函数，采用负采样(Negative Sampling)的方法以化简求解运算[16]。

(3)自适应K-means聚类

本文采取自适应K-means聚类的方法，对T-WV这一包含主题词排序和相关性信息的结构化数据矩阵进行聚类分析，从而达到主题最优个数判定的目的。传统K-means方法是解决聚类问题的经典方法之一，在这种算法中每个类用该类中对象的平均值来表示，该方法由MacQueen[17]首先提出，并由于其良好的运算性能而得到广泛的推广和应用。其基本步骤为[17]：

①选取k个样品作为初始凝聚点，或者将k个样品分成k个初始类，然后将这k个类的重心(均值)作为初始凝聚点；

②对除凝聚点之外的所有样品逐个归类，将每个样品归入凝聚点离它最近的那个类(通常采用余弦距离)，该类的凝聚点更新为这一类目前的均值，直至所有样品都归类完成；

③重复步骤②，直至所有的样品都不能再分配为止。

这种方法属于基于划分的聚类算法，其特点在于算法复杂度低、运算速度较快，因而非常适合大规模语料的运算，缺点是不能解决聚类个数选择问题，

一般均基于人为设定，无法排除主观因素。而本文提出的自适应聚类方法就是将聚类效果评价指标纳入目标函数、嵌入算法本身，通过迭代算法最终获取最优解，以达到确定主题最优个数的目的。

聚类个数的选择问题与聚类效果评判在狭义上具有相同的本质，有关聚类结果评价在学术界主要从两个维度展开：第一类是外部准则。用事先标记好的聚类结果来评价聚类效果。这类方法的问题在于事先标记结果并不能很准确地把握聚类意图，从而造成聚类评价的偏差；第二类是内部准则，将参与聚类的样本(在此为n个文本向量)作为评价聚类效果的主体，结合类间离差平方和与类内的离差平方两种因素对聚类个数进行干预和优化。其优点在于：内部准则不需要人为标签的事先设定，判别方法符合聚类的基本原则和思想，实现过程相对简单，适合本文的文本语料也易于拓展到更大规模文本语料中。因此，选用内部准则指标作为目标函数，进而确定主题个数的最优解。

4 实证分析

4.1 数据说明与处理

(1)数据来源

实验数据来源于CNKI，科技文献类型为有关“LDA主题模型”的研究文献[②]，所获取的文本信息主要包括题名、作者、摘要、关键词、文献来源和发表时间等科学文献文本语料的主要信息。数据的起止时间点为2013年10月14日～2017年1月22日，共有文献675篇，其中2013年之前的文献数量不足100，说明国内学者对这一问题的研究起步较晚。从2014年开始，有关该主题的文献研究数量超过100并逐年增多，表现为2014年138篇、2015年167篇、2016年177篇，2017年2篇(截至统计日期)。

(2)数据预处理

在删除个别信息不全的样本后，对原始数据进行清洗。首先提取关键词信息，构建关键词字典。考虑到文章中频繁出现一些术语，这些词汇的识别并不能完全依靠普通分词软件的常规词库进行，因此需要导入特定的关键词字典，便于进一步的数据分析工作。其次，提取摘要、题目这两部分最能反映文献核心价值的内容，采用含有关键词词典信息的jieba分词[③]工具，对这些语料进行分词与去除停用词处理。

采用CNKI内部的检索系统，可以自主生成电子表格形式的文本数据列表，将其转化成UTF-8编码形式，以便与Python系统兼容进行下一步的操作。原始文本数据样式如图2所示，其中包含作者、题名、关键词、摘要等反映文本核心内容的重要信息。

	A	B	C	D	E	F	G	H	I	J	K	L
1	作者	题名	文献来源	年	关键词	机构	摘要					
2	魏晨;张书	基于半监	山东大学		2011 姿态判别	西北工业	针对无监督的主题模型无法对图像主题进行类别标记、有监					
3	詹帆翠;	类别的聚	小型微型		2011 潜在狄里	江西财经	提出一种基于类别的聚的主题模型用于实现场景分类。不同					
4	王英帅;刘	一种基于	计算机应		2011 人名消歧	苏州大学	针对人物搜索的核心问题Web人名消歧进行了研究，根据同					
5	徐倪;陈世	利用特征	哈尔滨工		2011 场景分类	武汉大学	为了提高遥感图像场景分类精度，提出了一种基于增广LDA(					
6	张明警;基	于LDA计	计算机应		2011 自动文摘	苏州大学	近年来概率主题模型受到了研究者的广泛关注，LDA(Latent					
7	徐戈;王厚	自然语言	计算机学		2011 自然语言	北京大学	主题模型在自然语言处理领域受到了越来越多的关注。在该					
8	黄小亮;基	于LDA计	计算机工		2011 软件缺陷	复旦大学	传统的基于向量空间模型的软件缺陷分派方法，由于存在特					
9	骆翼林;	基于主题	计算机仿		2011 图像分割	华峰学院	研究文字图像中不同区域的分割问题。由于文字图像结构较					
10	谢昊;江红	一种面向	华东师范		2013 新浪微博	华东师范	随着新浪微博用户的不断增长，微博网站成为很多人获取信					
11	王振振;基	于LDA计	计算机科		2013 主题模型	北京工业	LDA(Latent Dirichlet Allocation)模型是近年来提出的					
12	陈铭夏;	基于LDA模	软件		2013 模式识别	北京邮电	当今的社交网络拥有庞大的用户数量和隐之而来的海量信息					
13	娄晓伟;基	于主题计	计算机研		2013 微博客;主	清华大学	近年来，以Twitter和新浪微博为代表的微博客正在世界范围					
14	毕娟;基于	频率	电子科技		2014 层次化主	电子科技	针对传统的社区发现算法大多基于网络拓扑结构寻找独立的					
15	卢文羊;基	于LDA计	山东大学		2014 社会网络	南京大学	针对传统社会网络链接预测方法忽视节点文本内容的问题，					
16	李然;张书	基于主题	计算机科		2014 摘要;LDA	北京理工	提出了一种基于LDA模型以及信息熵的文档自动摘要技术，基					
17	陶永才;基	于加权	计算机应		2014 加权;主题	郑州大学	针对微博信息量大、用户兴趣随时间变化特征，提出一种基					
18	王文辉;一	种面向	计算机工		2014 微博;Bio	中国科学	随着微博的日趋流行，新浪微博已成为公众获取和传播信息					
19	赵绍光;主	题模型	现代图书		2014 主题模型	中国医学	【目的】对基于主题模型的演化方法进行梳理与分析，总结					
20	李锦刚;基	于LDA-计	计算机应		2015 文本分类	合肥工业	SVM分类算法处理高维数据具有较大优势，但其未考虑语义的					
21	李湘东;引	采用LDA	图书馆论		2015 LDA模型;	武汉大学	探索对多种类型文献进行混合分类组织时LDA主题模型的可					
22	郭乔进;对	一种基于	模式识别		2014 雷兴鲍区	南京大学	雷兴鲍区域(ROI)的分类是医学图像的计算机辅助诊断过程					
23	龚严慧;基	于在线	计算机学		2013 潜在狄利	苏州大学	主题追踪因可以有效地汇集和组织分散在不同时间、地点的					
24	张伟男;对	一种面向	计算机学		2015 社区型问	哈尔滨工	基于统计机器翻译模型的问句检索模型。其相关性排序机制					
25	郑伸斌;基	于轨迹	计算机应		2015 视频分析	江苏省输	基于目标轨迹的异常行为检测算法忽略了轨迹内部信息，容					
26	王宇;基于	层次	图书情报		2014 主题发现	武汉大学	自动挖掘科技文献主题并识别主题变化对于科研工作者及时					
27	尚燕敏;利	融合链接	通信学报		2015 社交网络	中国科学	提出一种新的朋友推荐方法，该方法同时使用用户兴趣和朋					
28	王博;刘名	基于LDA	科研管理		2015 主题模型	大连理工	主题模型是一种有效提取大规模文本隐含主题的建模方法，					
29	朱俊鹏;基	于LDA主	数据采集		2015 LDA主题模	大理学院	借助mRNAs分析MicroRNAs(miRNAs)的研究已经用于阐述mi					
30	马宇峰;基	于LDA计	山东大学		2015 实体集合	华东理工	实体集合扩展是指给定某类别下若干示例作为种子，扩展得					
31	杜永萍;基	于主题	北京工业		2016 推荐系统	北京工业	针对文献推荐问题，提出了一种基于主题效能的学术文献推					

图2 原始文本数据样式

对于关键词，去除文字符号等冗余信息后将其作为字典备用。因为普通的切词工具无法对专业领域的词汇进行很好的切分，必须借助对该领域内容的重新学习达到精准的切分效果，而关键词提供了专业领域丰富的科技词汇，通过载入关键词词典，可以大幅提高分词的精准性。因此，在纳入关键词信息后，对摘要进行切词，这一过程同时加入停用词词典，可以达到数据清洗与降维的双重目的。某一篇摘要切词前后的对比结果，如图3所示。

摘要	切词摘要
针对无监督的主题模型无法对图像主题进行类别标记、有监督主题模型中类别信息的标记繁琐且受主观因素影响的问题，提出了一种半监督主题模型。提取图像中与位置无关的局部特征，用尺度不变特征变换对特征进行描述，用词袋模型将人脸图像表示成一组视觉单词的集合；在基于隐含狄利克雷分配(latent Dirichlet allocation，LDA)方法中的主题-单词层分布上引入少量的类别标记指导未标记样本的分类的基础上提出半监督隐含狄利克雷分配方法。在多姿态人脸判别任务上的测试结果表明该算法比无监督LDA算法分类率高9.0%~24.7%；对于部分遮挡人脸图像、未对齐的人脸图像的分类率比多姿态主成分分析法分别提高8.8%和21.5%~39.8%。结果表明该方法在少量样本标记的情况下，性能逼近有监督的隐含狄利克雷分配方法，且适用于其它图像分类问题。	针对 无监督 主题模型 无法 图像 主题 进行 类别 标记 监督主题模型 中 类别 信息 标记 繁琐 受 主观因素 影响 问题 提出 一种 监督主题模型 提取 图像 中 位置 无关 局部 特征 尺度 不变 特征 变换 特征 进行 描述 词袋模型 人脸 图像 表示 成 一组 视觉单词 集合 隐含狄利克雷分配 latent Dirichlet allocation LDA 方法 中 主题 单词 层 分布 引入 少量 类别 标记 指导 未 标记 样本 分类 基础 提出 监督 隐含狄利克雷分配 方法 姿态 人脸 判别 任务 测试 结果表明 算法 无监督 LDA算法 分类 率高 9.0 24.7 部分 遮挡 人脸 图像 未 对齐 人脸 图像 分类 率 姿态 主成分分析 法 分别 提高 8.8 21.5 39.8 结果表明 方法 少量 样本 标记 情况 下 性能 逼近 监督 隐含狄利克雷分配 方法 适用 图像分类 问题

图3 文本语料切词前后对比

由图3可见，数据预处理阶段加入关键词词典和传统的去停用词词典后，将非结构化的文本数据进行切分，同时达到数据清洗与降维的目的，为下一步LDA模型中T-W矩阵的获取提供了良好的数据基础。

本文硬件实验环境是一台Inter(R)Core(TM)i5-3470CPU、主频3.20GHz、内存4.00GB的PC，搭载Windows 10旗舰版的64位操作系统。软件则选用Python2.7，对于LDA模型、Word2Vec模型等理论模型通过调用Gensim包进行操作，并利用Python语言编程实现自适应聚类工作。

4.2 基于自适应聚类的K-wrLDA模型

(1)T-W矩阵的获取

要在此阶段完成对T-W的获取，首先要进行超参数选择从而实现模型。在主题模型中，有关两个超参数α和β的设定对模型效果有至关重要的影响。一般情况下，对超参数α选择是根据主题数目的变化而变化[18]，$\alpha=50/$主题数K，而β的选择基本固定为β=0.01。表1显示了不同主题数下α的取值。

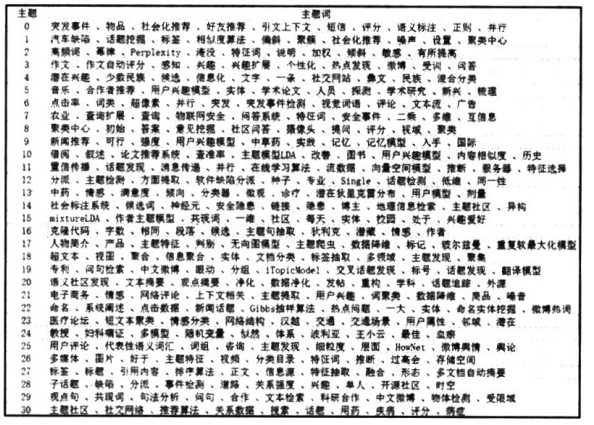
表1 不同主题数下α的取值

K	10	20	50	100	200	300	400	500
α	5	25	1	0.5	0.25	0.17	0.13	0.1

由此可见，超参数α的取值与主题个数成反比。在本文的实证分析中，α取值根据K的不同而浮动，以取得最佳的超参数设置值，β则统一设置为0.01。

在完成数据预处理以及超参数选择的基础上，进一步为文本数据构造LDA模型所需要的dic(词典)，并采用TF-IDF完成文本数据的初步数量化。为了选择最优主题数目，在这个阶段选择尽可能多的主题数备用和遴选，可以为后续主题数目的选取提供较大的空间，也可以使比较新旧两种模型在不同主题数目下的表现更具说服力。令K=100、iterations=2000，进行模型训练，每个主题内的主题词根据其概率的大小排序，说明排序靠前的主题词被划分在这一主题中的概率相对较大，反之亦然。在此取每个主题的前10个主题词，这些都是以较大概率留在主题内部，以反映主题内容的词汇，具有很强的代表性。由此所得到T-W矩阵。

图4展示了传统LDA模型下前30个主题及每个主题下所属概率最高的前10个主题词，它们共同构成一个30×10维的T-W矩阵。但由于主题内部的词具有独立性这一强假设，削弱了词语之间相关性甚至是主题潜在语义的刻画。因此，为了词语之间的相互关系能够得到体现，并能够将词语量化为词向量的形式(便于计算)，则需要在原始语料下引入词嵌入模型进行训练。



主题	主题词
0	突发事件 、物品 、社会化推荐 、好友推荐 、引文上下文 、短信 、评分 、语义标注 、正则 、并行
1	汽车轨迹 、话题挖掘 、标签 、相似度算法 、偏斜 、聚簇 、社会化推荐 、噪声 、设置 、聚类中心
2	高频词 、幂律 、Perplexity 、海投 、特征词 、说明 、加权 、螺料 、敏感 、有所提高
3	作文 、作文自动评分 、感知 、兴趣 、兴趣扩展 、个性化 、热点发现 、微博 、受词 、问答
4	潜在兴趣 、少数民族 、候选 、信息化 、文学 、一条 、社交网站 、象文 、民族 、混合分类
5	音乐 、合作者推荐 、用户兴趣模型 、实体 、学术论文 、人员 、探测 、学术研究 、新兴 、梳理
6	点击率 、词类 、相像素 、并行 、突发 、突发事件检测 、视觉词语 、评论 、文本流 、广告
7	农业 、查询扩展 、查询 、物联网安全 、问答系统 、特征词 、安全事件 、二维 、多维 、互信息
8	聚类中心 、初始 、答案 、意见挖掘 、社区问答 、摄像头 、提问 、评分 、视域 、聚类
9	新闻推荐 、可行 、强度 、用户兴趣模型 、中草药 、实践 、记忆 、记忆模型 、人手 、国际
10	借阅 、叙述 、论文推荐系统 、查准率 、主题模型LDA 、改善 、图书 、用户兴趣模型 、内容相似度 、历史
11	置信传播 、话题发现 、消息传递 、并行 、在线学习算法 、资查据 、向量空间模型 、推断 、服务器 、特征选择
12	分词 、主题检测 、方面提取 、软件缺陷分派 、种子 、专业 、Single 、话题检测 、低维 、同一性
13	中药 、情感 、测量度 、倾向 、分类器 、锻炼 、治疗 、潜在狄里克雷分布 、用户模型 、测量
14	社会标注系统 、候选词 、神经元 、安全隐患 、链接 、隐患 、博主 、地理信息检索 、主题社区 、异构
15	mixturedLDA 、作者主题模型 、共现词 、一维 、社区 、每天 、实体 、挖掘 、处于 、兴趣爱好
16	克隆代码 、字数 、相同 、段落 、候选 、主题句抽取 、状判克 、潜藏 、情感 、作者
17	人物简介 、产品 、主题特征 、判别 、无向图模型 、主题爬虫 、数据降维 、标记 、锁尔兹曼 、重复软最大化模型
18	短文本 、视图 、聚合 、信息聚合 、实体 、文档分类 、标签抽取 、多视域 、主题发现 、聚集
19	专利 、词句检索 、中文微博 、眼动 、分组 、iTopicNodel 、交叉话题发现 、标号 、话题发现 、翻译模型
20	语义社区发现 、文本摘要 、要点概要 、净化 、数据净化 、发帖 、重构 、学科 、话题追踪 、外源
21	电子商务 、情感 、网络评论 、上下文相关 、主题模型 、用户兴趣 、词聚类 、数据降维 、商品 、噪音
22	命名 、系统阐述 、点击数据 、新闻话题 、Gibbs抽样算法 、热点问题 、一大 、实体 、命名实体挖掘 、微博热词
23	医疗论坛 、短文本聚类 、情感分类 、网络结构 、反超 、交通 、交通场景 、用户属性 、邻域 、潜在
24	教授 、妇科病证 、多模型 、随机变量 、知然 、体系 、读判定 、王小云 、最佳 、监督
25	用户评论 、代表性语义词汇 、词组 、咨询 、主题发现 、细粒度 、壁面 、HowNet 、微博舆情 、舆论
26	多媒体 、图片 、好于 、主题特征 、视频 、分类目录 、特征词 、推断 、过高会 、存储空间
27	标签 、标题 、引用内容 、排序算法 、正文 、信息源 、特征抽取 、融合 、形态 、多文档自动摘要
28	子话题 、缺陷 、分类 、事件检测 、道路 、关系强度 、兴趣 、单人 、开源社区 、时空
29	视点句 、共现词 、句法分析 、问句 、合作 、文本检索 、科研合作 、中文微博 、物体检测 、受限域
30	主题社区 、社交网络 、推荐算法 、关系数据 、搜索 、话题 、用药 、疾病 、评分 、病症

图4 T-W矩阵(部分)

(2)T-WV矩阵的数量化构造

在得到文本形式下的T-W矩阵之后，要将其转化为含有词语间相关信息的数量化表示。传统的词向量表达方法是One-Hot Representation，这种方法所构造的词向量长度就是整个词典的大小，高维且稀疏，无法很好地表达词语之间的语义信息。因此本文采用Distributed Representation方式构造的词向量，这种方法可以将词映射到一个低维、稠密的实数向量空间(本文选取的空间大小是64维)，词义相近的词能够在数学意义的空间距离上得到数量化的体现。例如“西红柿”和“番茄”“土豆”和“马铃薯”这类词汇原本具有同样的语义，但在传统词向量表示方法中计算，会认为其相似度为0；相反“苹果”可能表示水果，也有可能表示某一电子类产品品牌，两者在不同的语境下的语义截然不同，由于传统模型都遵从词袋假设，因此会认为它们之间的相似度为1。而Word2Vec模型就是采用Distributed Representation的表达方式，其优点在于它可以结合上下文和语境，对词义有更准确的判断。因此，采用这种词向量表达方法可以有效解决LDA模型中T-W矩阵的词语

间相似性的识别与表达问题。

将关键词、摘要、标题这三个最能反映文章核心技术与内容的部分作为训练集，对Word2Vec模型进行训练。训练完成的模型，可以认为其充分识别和掌握了原始语料内部的词汇相关度信息，故其词向量的数值结果具有较高的信息价值。为了检测Word2Vec模型的训练结果，将“LDA”作为待测词输入到该模型中，Word2Vec模型会根据事先训练好的模型给出待测词的相关词汇以及相关系数，反馈出与此词相关度较高的词汇。在此设置相关词汇的阈值=6，Word2Vec模型就会反馈与待测词汇最为相关的6个词语，结果展示如表2所示。

表2 基于Word2Vec模型的“LDA”相关词汇

相似词	相似度
LDA模型	0.743
主题模型	0.722
概率主题模型	0.707
潜在狄里克雷	0.599
LDA算法	0.556
生成模型	0.512

表2反映了LDA的相关词汇，相似度计算结果显示前三个词汇与测试词的相关性极高，均大于0.7，而前6个词的相似度均达到0.5以上，体现出Word2Vec模型优良的词语间相关性识别能力。说明Word2Vec模型根据原始语料，能够完成词语间相关性的量化与测量。因此，采用该方法对LDA模型的100个主题词进行词向量化处理，即对T-W矩阵进行词向量化表示。更加科学的词向量化，在精准刻画词义之间相似度的同时，为进一步的聚类分析提供更加可信的数据资料。

在具体操作过程中，将LDA模型中的T-W矩阵输入给Word2Vec模型，得到该矩阵词汇的量化结果。在得到所有语料的词向量后，鉴于词向量可加减的性质，原则上可以将其进行加法运算以压缩整个矩阵的维度。但考虑到传统T-W矩阵中的词语排序包含主题词的概率分布信息，倘若直接采用加法处理会造成该信息的损失。因此采用第二种方式，对每个主题的词向量进行拼接，构成既包含主题词概率分布信息，又包含词汇相似度概念的词汇长向量。此方法虽然造成语料维数的增加，但保留了LDA模型中语料的重要信息，并未对后续计算带来算法上时间完成度的困难，因此具有较强的可行性和科学性。与T-W矩阵不同的是T-WV矩阵已经将前者的汉语文本词汇矩阵转换为词向量矩阵，因此后者是一个映射在高维空间的数量化矩阵。以上只完成了模型的wrLDA模型部分，对于主题最优个数问题还未涉及，需要通过自适应聚类来解决，因此要进一步在Python2.7下进行基于T-WV数据框架的K-means聚类分析。

(3)自适应K-means聚类

传统LDA模型由于没有很好的主题个数选择标准，通常根据人为主观模式进行确定。针对新模型，本文在聚类过程中加入自适应迭代机制，解决模型的主题个数选择问题。其思路为将传统LDA模型主题个数的选择这一陌生问题转化为聚类分析中聚类个数选择问题。给定一个关于聚类效果的判别标准，通过计算机迭代计算，得出最优解。有关聚类个数选取问题具有相对成熟的技术体系，可选择的方法较多。大致可以分为外部指标和内部指标两种考核体系，前者需要对文本进行标记，这一过程缺乏效率，并且时常会出现标记本身与聚类目标不匹配的问题，而后者从聚类结果出发，根据类间与类内的差异性评判聚类效果的好坏，简单易行且具有显著的统计学意义。因此，本文着眼于聚类的内部性评价指标，选用伪F统计量进行判断。

设总文档数量为n，聚类时将所有文档合并成k个主题$G_1, G_2, \cdots, G_k$，主题G_i的文档数和重心分别是n_i和$\bar{x}_i$，i=1，2，…，k，则$\sum_{i=1}^{k} n_i=n$，所有文档的总重心为$\bar{x}=\frac{1}{n}\sum_{i=1}^{k} n_i\bar{x}_i$，令$W=\sum_{j=1}^{n}(x_j-\bar{x})'(x_j-\bar{x})$为所有文档的总离差平方和，$w_i=\sum_{j\in G_i}(x_j-\bar{x}_i)'(x_j-\bar{x}_i)$为主题$G_i$中文档的类内离差平方和，$P_k=\sum_{i=1}^{k} W_i$为k个主题内离差平方和之和。W可做如下分解[19]，如公式(6)所示。

$$\begin{aligned}W&=\sum_{j=1}^{n}(x_j-\bar{x})'(x_j-\bar{x})=\sum_{j=1}^{k}\sum_{j\in G_i}(x_j-\bar{x})'(x_j-\bar{x})\\&=\sum_{i=1}^{k}\sum_{j\in G_i}(x_j-\bar{x}_i+\bar{x}_i-\bar{x})'(x_j-\bar{x}_i+\bar{x}_i-\bar{x})\\&=\sum_{i=1}^{k}\Big[\sum_{j\in G_i}(x_j-\bar{x}_i)'(x_j-\bar{x}_i)+n_i(\bar{x}_i-\bar{x})'(\bar{x}_i-\bar{x})+\\&\quad 2\sum_{j\in G_i}(x_j-\bar{x}_i)'(x_i-\bar{x}_i)\Big]\\&=P_k+\sum_{i=1}^{k}n_i(\bar{x}_i-\bar{x})'(\bar{x}_i-\bar{x})\end{aligned}\qquad(6)$$

令 $R^2=1-Pk/W=\sum_{i=1}^{k}n_i(\bar{x}_i-\bar{x})'(\bar{x}_i-\bar{x})\Big/W$，构造伪F统计量[19]，如公式(7)所示。

$$伪F=\frac{(W-P_k)/(k-1)}{P_k/(n-k)}=\frac{n-k}{k-1}\cdot\frac{R^2}{1-R^2} \tag{7}$$

其中，分子表示主题间离差平方和，分母表示主题内离差平方和，$\frac{n-k}{k-1}$ 为调整系数亦称为惩罚项。伪F统计量公式由惩罚系数项和反映组间、组内差异项两部分构成，伪F统计量的值越大说明聚类效果越好，反之亦然。因此在新模型的自适应聚类阶段，本文将聚类判别中的伪F统计量应用到主题模型中，设置其目标函数如公式(8)所示。

$$\max\frac{n-k}{k-1}\cdot\frac{R^2}{1-R^2} \tag{8}$$

传统LDA假设服从词袋模型，认为词与词之间是相互独立的关系，但这一要求过于严苛，几乎不符合现实情况。因此，新模型克服了词袋模型的独立性假设，在原始语料上训练Word2Vec模型，增强了词语间相关性的识别能力，并采用自适应聚类方法，通过设置目标函数以及K的遍历范围，从而获得聚类的局部最优解，即为最优主题个数。整个模型的实现流程如下：

(1)对初始参数以及超参数赋值。利用传统LDA模型对文档集合D(L)进行计算，求出$(T-W)_{n\times k}$的分布矩阵，其中n表示主题中词的个数，k为主题个数，令n=10，k=100。

(2)将原始语料进行Word2Vec模型训练可以得到从微观角度考虑词语之间相关性的词向量，并通过LDA模型的T-W矩阵给出与之对应的词向量的数量化反馈，得到一个全新的有关词向量相关性和主题词排序信息的数量关系矩阵$(T-WV)_{u\times k}$，其中m为词向量的维度，u=m×n，本文令m=64，则u=640。

(3)将数量化矩阵T-WV纳入自适应K-means算法，通过余弦距离计算文档间的相似度。令k=1，…，100，遍历该阈值范围内的所有情形，自适应聚类算法会根据内嵌的评判目标函数伪F统计量给出聚类最佳个数。

为了让这一过程更加直观，记录k=1，…，100每种情形下目标函数的得分值，并生成关于聚类个数的得分曲线，如图5所示。该伪F统计量的曲线呈现先增长后下降的趋势，并且在k=33时，聚类效果最佳。因此，最优主题数目k值也由此确定，K-wrLDA模型得以完成。

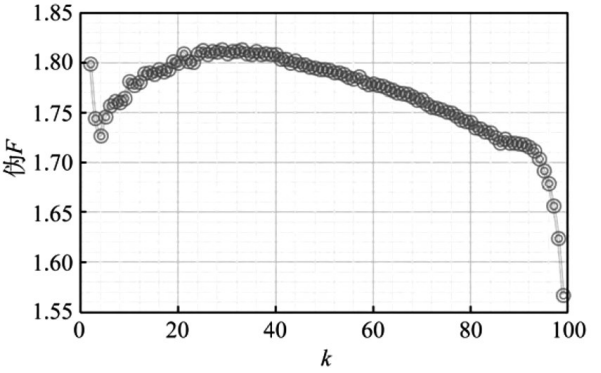


图5 伪F统计量趋势

4.3 模型评判

困惑度指标是LDA模型的原作者Blei等提出的一种反应模型泛化能力的指标，在评价模型的优劣上具有一定的代表性和普遍性，故本文也采用该指标进行模型评价。对于传统LDA模型，令其主题个数k选取范围为1～100，新模型主题个数选择范围与之相同，以便展开对比分析。为了评估新模型与传统LDA模型的优劣，逐一计算k取值范围内的困惑度指标得分并比较其趋势，从而完成新模型和传统模型的性能测试。所谓困惑度就是文档在划分主题时确定性的评判，反映的是模型对新样本的适用性。其计算如公式(9)[20]所示。

$$P(\tilde{W}|M)=\exp-\frac{\sum_{m=1}^{M}\log p(\tilde{w}_{\tilde{m}}|M)}{\sum_{m=1}^{M}N_m} \tag{9}$$

困惑度值越小，表示该模型对新样本的分类效果越好，泛化能力越强，反之亦然。其中M表示文本集中的文本数，N_m表示文档m的长度，$\log p(\tilde{w}_{\tilde{m}}|M)$为第m篇文档中词的概率值(所有词的概率乘积)，公式(9)的计算难点主要集中在分子部分，如公式(10)所示。

$$\begin{aligned}p(\tilde{w}_{\tilde{m}}|M)&=\prod_{n=1}^{N_{\tilde{m}}}\sum_{k=1}^{K}p(w_n=t|z_n=k)\cdot p(z_n=k|d=\tilde{m})\\&=\sum_{t=1}^{V}\left(\sum_{k=1}^{K}\varphi_{k,t}\cdot\vartheta_{\tilde{m},t}\right)^{n_{\tilde{m}}}\end{aligned} \tag{10}$$

其中，k表示主题个数，n表示文档中词的个数。$p(z_n=k|d=\tilde{m})=\vartheta_{\tilde{m},k}$为当前文本下主题为k的概率，$p(w_n=t|z_n=k)=\varphi_{k,t}$则表示主题为k时当前文本下第t个

词的概率。分子变得可解[19],如公式(11)所示。

$$\log p(\tilde{w}_{\tilde{m}}|M)=\sum_{t=1}^{V} n_{\tilde{m}}^{(t)} \log\left(\sum_{k=1}^{K} \varphi_{k,t}\cdot \vartheta_{\tilde{m},t}\right) \quad (11)$$

本文用75%的文档作为训练集以训练模型,25%的数据作为测试集计算困惑度,以反映模型的泛化能力。

根据上述思想,采用困惑度指标对新旧模型的泛化能力进行比较。k的取值为:1~100,编写代码遍历给定k阈值范围内的所有情况,曲线走势如图6所示。

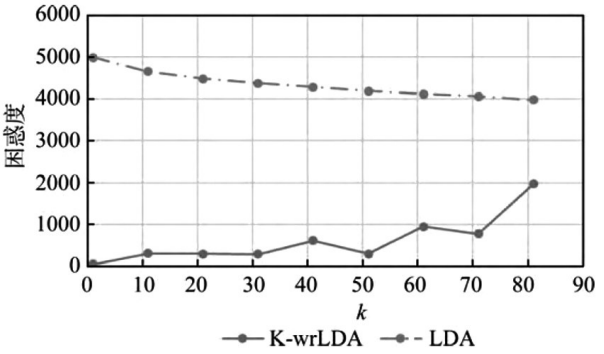


图6 基于困惑度的新旧模型性能评估

图6显示了不同主题聚类数目下的困惑度指标走向,其中实线为K-wrLDA模型的困惑度指数值,虚线为传统LDA模型的困惑度指数趋势。显然在k值的任何一种情况下,新模型的困惑度始终低于传统LDA模型,说明本文构造的K-wrLDA模型具有更优越的性能,故新模型具有一定的实用和推广价值。需要说明的是由于可供主题聚类的最大主题数为100,在主题数k=100时,两者的困惑度指标一致,因此随着主题数目增加,新模型的困惑度会出现上升趋势,但依然始终具有比传统模型好的表现。在实际应用过程中,主题选择太多会给主题划分和解释造成一定难度,选择太少又容易掩盖一些小主题,因此主题数的选择会根据语料体量的大小进行调整,一般集中在10~50之间。在这一范围内的新模型困惑度指数值非常低,并表现出较为平稳的走势,因此新模型不但克服了传统模型的词袋假设缺陷,且在模型的泛化能力方面具有更好的表现。

4.4 K-wrLDA模型与传统LDA模型主题识别对比分析

基于上述关于困惑度指标下的K-wrLDA模型与传统LDA模型的评价,发现前者在各个主题数目下均具有更加优异的表现。为了更直观地比较两种模型在主题识别方面的能力,采用上述语料,在获取最优主题个数的情形下,分别针对传统LDA模型和K-wrLDA模型进行实证分析。鉴于数据量及展示的问题,在此分别取前10个主题及其前10个主题词进行对比,如表3和表4所示。

为了便于展示与比较,在此选取33个主题中的前10个主题以及主题词中的前10个主题词进行对比。直观可看出,表3的主题之间相似性和交叉性较强,出现重复和同义词汇的概率较大,而单个主题内部之间的主题词语义不够集中,相对分散,对主题的诠释力度较差。例如主题0、主题1、主题3等均涉及情感分析方面,主题3、主题4、主题6和主题7是关于短文本评论方面的语料分析。可见这些主题之间的区分度不高,存在主题内容的重叠交叉,故认为其主

表3 传统LDA模型下的主题识别结果

主题0	主题1	主题2	主题3	主题4	主题5	主题6	主题7	主题8	主题9
情绪	情感	微博	评论	观点	专利	兴趣	词向量	学科	人物简介
新闻推荐	情感分类	推荐	投诉	评论	主题演化	专家	方剂	知识流	电子书
新闻	评论	用户	子话题	情感分析	在线	评分	点击率	克隆代码	子话题
句子	运动	短文本	信息增益	标注	期刊	评论	评分	文献	农业
interest	特征提取	微博用户	产品	观点挖掘	文本流	项目	遥感	分级	电影
读者	评论文本	推荐算法	翻译	软件	中医药	用户	提案	线程	输入
医疗论坛	实体	词汇	正文	合作	文本分割	偏好	主题模型可视化	问句检索	作者
消息传递算法	聚类	个性化推荐	分派	症状	年度	用户兴趣	伪相关反馈	聚类中心	情感
词语	监督	作文	主题分割	借阅	句子	信息检索	账号	情感摘要	查询推荐
Web服务	句子	协同过滤	情绪	临床	文献	模式	社会化推荐	主题抽取	日志

表4 K-wrLDA模型下的主题识别结果

主题0	主题1	主题2	主题3	主题4	主题5	主题6	主题7	主题8	主题9
评论	专利	问句检索	查询	医疗论坛	随机变量	新闻	视图	教育资源	文本分割
短文本分类	发明人	运动	分布式	舆论	超文本	推荐算法	低质量回帖	视觉单词	任务模型
点击率	投诉	广告投放	word2vec	脑血管病	情感分类	人群	关键词抽取	提案	语义信息
句子	汽车缺陷	实体	矩阵分解	话题检测	文档	用户兴趣	博客	主题模型可视化	数字资源
相似性度量	遥感	关联主题	词聚类	查询	信息熵	用户评论	安全隐患	观点	特征项
词向量	作弊	单机	共享内存	咨询	网络舆情	粒计算	交通	视频	词向量
观点摘要	mixtureLDA	词项	文本建模	语义指纹	主题情感混合模型	online	关键词集	帐号	投放
朴素贝叶斯	词义	投放	消息传递算法	标记	自动应答系统	个性化推荐	隐患	句群	主题特征
引文上下文	用户	相似度算法	线程	文章	标签抽取	新浪微博	查询	语义标注	偏斜
共享主题	兴趣	热点话题	数字	相似矩阵	马尔科夫	调控	句法分析	标注单词	阅读概率

题识别效果不理想。从内部主题词来看，词汇的指向性相对分散，以主题9为例，其主题词包含电子书、农业、电影、日志等多范围的词汇，难以从中对该主题进行主题聚焦。而这些现象在表4中得到不同程度的改善。各个主题之间的区分度较为明显，大致可以分为：短文本评论、专利文本分析、广告投放效果、语料查询、医疗话题检测、情感分析、新闻追踪、可视化、教育、文本分割。并且各个主题内部的主题词交叉性较低，较表3而言，具有更强的凝聚性。以主题4为例，主题词包括文本处理的部分术语，其他均针对医疗方面的内容展开。因此，K-wrLDA模型不但能够自主确定最优主题个数，并且在主题识别方面较传统LDA模型而言具有一定的优势。

5 结论

本文旨在进一步优化LDA模型，并提出一种确定最优主题数的方法，使其更好地服务于文本数据的分析与处理。在传统LDA模型的基础上加入深度学习层面的Word2Vec模型，加强主题词之间相似性关系的刻画，并以此为数据基础进行主题自适应聚类。通过困惑度指标对新旧模型进行性能评估，发现在任何一种取值下新模型的性能均优于传统LDA模型。并将一个陌生问题回归到具有相对成熟理论支撑的统计学问题当中，从而有效解决了主题个数的选择问题。局限在于聚类方法的适用性和算法复杂度之间的矛盾，因此未来研究方向即为结合多元统计分析理论突破传统的聚类手段，试图将兼顾方法和效率的聚类技术应用到主题数目选择当中。

实验数据取自CNKI数据库中有关LDA主题模型的科技文献，在针对该项语料的实验过程，均体现出新模型的优越性能。本文并没有对其他类型的文本类型，例如新闻、社交媒体的短文本等进行验证，但上述工作内容与技术手段均可以在其他大规模文本语料中进行对接，因此本文提出的模型为处理其他类型的文本语料提供了一定的借鉴思路。此外，自适应聚类的结果囿于阈值范围，仅是局部最优解，但考虑到聚类分析的实际效果和需求，阈值远远高于实际划分个数的可能性，因此具有较强的说服力。

注释：

①在这一过程中还可以得到D-T矩阵，而本文研究并没有涉及该矩阵的应用，故视为模型的副产品。

②考虑到线性判别分析(Linear Discriminant Analysis)模型的缩写也为LDA，故在使用检索系统时，在高级搜索中加入主题模型这一约束条件，有效杜绝了线性判别分析类文献的混入，从而保证了文本数据搜集的精准性。

③https://pypi.python.org/pypi/jieba/。

参考文献：

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

[2]Blei D M, Lafferty J D. Correlated Topic Models[J]. Advances in Neural Information Processing Systems, 2005, 18: 113-120.

[3]关鹏，王曰芬，傅柱.不同语料下基于LDA主题模型的

科学文献主题抽取效果分析[J].图书情报工作,2016,60(2):112-121.

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

[5]石晶,胡明,石鑫,等.基于LDA模型的文本分割[J].计算机学报,2008,31(10):1865-1873.

[6]Hajjem M, Latiri C. Combining IR and LDA Topic Modeling for Filtering Microblogs[J]. Procedia Computer Science, 2017, 112: 761-770.

[7]廖列法,勒孚刚,朱亚兰.LDA模型在专利文本分类中的应用[J].现代情报,2017,37(3):35-39.

[8]刘江华.一种基于kmeans聚类算法和LDA主题模型的文本检索方法及有效性验证[J].情报科学,2017,35(2):16-21.

[9]Teh Y, Jordan M, Beal M, et al. Hierarchical Dirichlet Processes[J]. Journal of the American Statistical Association, 2007, 101(476): 1566-1581.

[10]颜端武,陶志恒,李兰彬.一种基于HDP模型的主题文献自动推荐方法及应用研究[J].情报理论与实践,2016,39(1):128-132.

[11]唐浩浩,王波,席耀一,等.基于HDP的无监督微博情感倾向性分析[J].信息工程大学学报,2015,16(4):463-469.

[12]曹娟,张勇东,李锦涛,等.一种基于密度的自适应最优LDA模型选择方法[J].计算机学报,2008,31(10):1780-1787.

[13]关鹏,王曰芬.科技情报分析中LDA主题模型最优主题数确定方法研究[J].现代图书情报技术,2016(9):42-49.

[14]茆诗松,王静龙,濮晓龙.高等数理统计[M].北京:高等教育出版社,2006:446-449.

[15]Hinton G E. Learning Distributed Representations of Concepts[C]//Proceedings of the 8th Annual Conference of the Cognitive Science Society, 1986.

[16]Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality [C]//Proceedings of the Neural Information Processing Systems Conference, 2013.

[17]MacQueen J. Some Methods for Classification and Analysis of Multivariate Observations[C]//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967.

[18]Wei X, Croft W B. LDA-based Document Models for Ad-Hoc Retrieval[C]//Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2006: 178-185.

[19]王学民.应用多元分析[M].上海:上海财经大学出版社,2003:217-218.

[20]Heinrich G. Parameter Estimation for Text Analysis[R]. vsonix GmbH+University of Leipzig, 2008: 29-30.

Optimizing LDA Model with Various Topic Numbers: Case Study of Scientific Literature

Wang Tingting Han Man Wang Yu

Abstract: [Objective]This paper proposes a K-wrLDA model based on adaptive clustering, aiming to improve the subject recognition ability of traditional LDA model, and identify the optimal number of selected topics.[Methods] First, we used the LDA and word2vec models to construct the T-WV matrix containing the probability information and the semantic relevance of the subject words. Then, we selected the number of topics based on the evaluation of clustering effects and the pseudo-F statistic. Finally, we compared the topic identification results of the proposed model with the old ones.[Results]The optimal number of topics was 33 for the proposed model, which also has lower level of perplexity than the traditional ones.[Limitations]The sample size needs to be expanded.[Conclusions]The proposed model, which has better recognition rate than the traditional LDA model, could also calculate the optimal number of topics. The new model may be applied to process large corpus in various fields.

Key words: Topic model; Word embedding; Adaptive clustering; Perplexity

