时间加权的 TF-LDA 学术文献摘要主题分析

伍 哲,杨 芳
(西安邮电大学 计算机学院,陕西 西安 710121)

摘 要:随着网络的发展,主题提取的应用越来越广泛,尤其是学术文献的主题提取。尽管学术文献摘要是短文本,但其具有高维性的特点导致文本主题模型难以处理,其时效性的特点致使主题挖掘时容易忽略时间因素,造成主题分布不均、不明确。针对此类问题,提出一种基于 TTF-LDA(time+tf-idf+latent Dirichlet allocation)的学术文献摘要主题聚类模型。通过引入 TF-IDF 特征提取的方法,对摘要进行特征词的提取,能有效降低 LDA 模型的输入文本维度,融合学术文献的发表时间因素,建立时间窗口,限定学术文献主题分析的时间,并通过文献的发表时间增加特征词的时间权重,使用特征词的时间权重之和协同主题引导特征词词库作为 LDA 的影响因子。通过在爬虫爬取的数据集上进行实验,与标准的 LDA 和 MVC-LDA 相比,在选取相同的主题数的情况下,模型的混乱程度更低,主题与主题之间的区分度更高,更符合学术文献本身的特点。

关键词:LDA;主题模型;学术文献;TF-IDF;时间因素
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2020)01-0194-07
doi:10.3969/j.issn.1673-629X.2020.01.035

A Thematic Analysis Method of Academic Documents Based on TF-IDF and LDA

WU Zhe, YANG Fang
(School of Computer Science, Xi'an University of Posts and Telecommunications, Xi'an 710121, China)

Abstract: With the development of network, topic extraction has been applied more and more widely, especially in academic literature. Although abstracts of academic literature are short texts, their high dimensionality makes it difficult to deal with text topic models, and their timeliness makes it easy to ignore the time factor in topic mining, resulting in uneven and unclear topic distribution. In order to solve these problems, a topic clustering model of academic literature abstracts based on TTF-LDA (tf-idf+latent Dirichlet allocation) is proposed. By introducing TF-IDF feature extraction method to extract feature words from abstracts, the extraction of feature words in the abstract can effectively reduce the input text dimension of LDA model, integrate the publication time factor of academic literature, establish a time window, and limit the time of subject analysis of academic literature. The time weights of feature words are increased by the publication time of documents, and the time weights of feature words are combined with the collaborative topics to guide the feature lexicon as the influencing factors of LDA. Through experiments on data sets crawled by crawlers, compared with standard LDA and MVC-LDA, the chaotic degree of the model is lower when the number of topics is the same, and the distinction between topics is higher, which is more in line with the characteristics of academic literature itself.

Key words: LDA; thematic model; academic literature; TF-IDF; time factor

0 引 言

学术文献是一种特殊的记录,或者可以称之为科学的总结,记录一种学术课题的新的科研成果,也总结一些创新性的见解。思路是应用某种已知的原理,对实际问题进行解决的进程叙述,可用来与其他人进行交流,多在学术性的会议上进行宣读,进行讨论,多数发表于相应领域的刊物上,其他则作为别的用途的书面文件[1]。文献是一种载体,用来传播学术性知识,人们通常阅读文献来获取知识,其可以反映人们在一定社会历史阶段的知识水平,其更是科学研究的基础[2]。随着社会的发展,文献的种类和数量越来越多,相关的研究人员在从事一项科研之前,需要进行准备工作,包

收稿日期:2019-01-25 **修回日期**:2019-05-28 **网络出版时间**:2019-09-24
基金项目:陕西省教育专项科研计划项目(15JK1679);西安市科技创新引导项目(201805040YD18CG24(7))
作者简介:伍 哲(1994-),男,硕士,研究方向为数据挖掘、舆情分析。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190924.1537.048.html

括获得这项科研的相关知识,进行人员分配,设计科研的实施方法,以及定期目标等,在积累基础的阶段,目前仍然还得阅读大量的学术文献,了解这项科研的全面知识,并且得到这项研究所属学术领域的最新研究热点。数量如此庞大的学术文献,人工进行分析显然速度很慢,无法达到目前社会的效率要求。搜索引擎是处理这一问题的工具之一,但其只能帮助科研人员筛选出符合检索条件的文章列表,这些列表对于科学研究需要的主题没有什么实用性价值,科研人员仍然需要通过大量阅读来熟知这些列表的内容,这需要付出很多时间和精力。如何更加有效地快速得到海量专业学术文献主题信息,更加直观地得到学术文献主题的结果信息,使科研人员迅速了解学术文献的热点和发展,判断该学术领域的发展方向,从而快速进行下一项任务。显然,减少人工查看分析时间,节省科研人员的精力,是一个急需解决的现实问题。

因此,为了能够高效、准确地提取学术文献的主题,提出一种 TF-IDF[3] 结合 LDA 的学术文献主题分析方法。该方法采用分词和停用词词典对文献集进行预处理,使用 TF-IDF 对其进行特征提取,降低维度,使用特征词构建主题引导特征词词库引导主题的生成,并加入时间因素,提出时间权重,综合特征词权重和时间权重计算总的影响权重,引导主题的概率分布,最后采用 LDA 主题模型得到主题分布情况。

1　相关介绍

1.1　TF-IDF

TF-IDF 是一种用于信息检索与数据挖掘的常用加权技术[4]。TF 意思是词频,指的是某一个给定的词语在该文件中出现的频率。IDF 意思是逆文本频率,在 IDF 中,词的集合中的一个词,有这样的特点,相对于其他的词,这个词在文档集中很少出现,但这个词在某一篇文档中却经常出现。显然这个词对于整个文档集而言没有任何意义,不是整个文档集的关键词,但对于这篇文章来说很重要,这个词就是这篇文章的关键词。那么怎样用一个指标来表示这种特性,如何去衡量这个词,怎么给它一个相对这篇文章较高的,而相对总体文档集没有作用的权重呢,这是一个问题。通常,这个特性是一个具有调整功能的变量,则需要定义一个重要性的调整系数来解决这个问题,用统计学语言表达就是在词频统计的基础上,对每个词项分配一个“重要性”的调整系数,这个词的出现次数和它的权重呈反比,出现的多反而权重小,出现的少反而权重大,具有重要作用,这就是通常所说的逆文档频率。综上,TF-IDF 的主要思想如下:对于某个属于词集合的词,如果在一篇文章中出现的频率(TF)高,并且在其他文章中很少出现(IDF),则认为此词是这篇文章的关键词,即特征词,与其他词相比,具有代表性,有很好的类别区分能力,能代表这篇文章。这个算法的细节如下:

词频(TF)= 词在文档中的出现次数/该文档中所有字词的出现次数之和

逆向文件频率(IDF)= log(文档总数/包含该词语的文档数目)

TF-IDT = TF * IDF

1.2　LDA

LDA(隐含狄利克雷分布)是目前一种比较主流的主题模型,也是一种典型的词袋模型[5]。它是一种非监督机器学习技术,可以展现离散型数据集的概率增长,具有三层,分别为文档集层、主题层及特征词层,每层均由相应的随机变量或参数控制。它可以将文档集合中的每篇文档的主体以概率分布的形式给出,从而分析一些文档抽取出它们的主题分布,然后可以根据主题进行文本分类或者是主题聚类。LDA 采用贝叶斯估计的方法,假设文档的主题分布和主题的特征词分布的先验分布都是 Dirichlet 分布(狄利克雷分布),认为所有的文档存在 K 个隐含主题,要生成一篇文档,首先生成该文档的一个主题分布,然后再生成词的集合;要生成一个词,需要根据文档的主题分布随机选择一个主题,然后根据主题中词的分布随机选择一个词,重复这个过程直至生成文档。

LDA 是一种使用联合分布计算在给定观测变量下隐藏变量的条件分布(后验分布)的概率模型,观测变量为词的集合,隐含变量为主题[6]。LDA 的生成过程对应的观测变量和隐藏变量的联合分布如式 1 所示:

$$p(\beta_{1:K},\theta_{1:D},Z_{1:D},W_{1:D}) = G * H * J \tag{1}$$

$$G = \prod_{i=1}^{K} p(\beta)$$

$$H = \prod_{d=1}^{D} p(\theta_d)$$

$$J = \left(\prod_{n=1}^{N} p(Z_{d,n} \mid \beta_d) p(W_{d,n} \mid \beta_{1:K},(Z_{d,n}))\right)$$

其中,β 表示主题,θ 表示主题的概率,Z 表示特定文档或词语的主题,W 为词语。$\beta_{1:K}$ 为全体主题集合,其中 β_k 是第 k 个主题的词的分布。第 d 个文档中该主题所占的比例为 θ_d,其中 $\theta_{d,k}$ 表示第 k 个主题在第 d 个文档中的比例。第 d 个文档的主题全体为 Z_d,其中 $Z_{d,n}$ 是第 d 个文档中第 n 个词的主题。第 d 个文档中所有词记为 W_d,其中 $W_{d,n}$ 是第 d 个文档中第 n 个词,每个词都是固定的词汇表中的元素。$p(\beta)$ 表示从主题集合中选取了一个特定主题,$p(\theta_d)$ 表示该主题在特定文档中的概率,大括号的前半部分是该主题确定时该文档第 n 个词的主题,后半部分是该文档第 n

个词的主题与该词的联合分布。连乘符号描述了随机变量的依赖性,用概率图模型表述如图 1 所示。

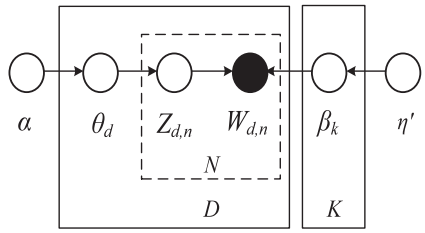

图 1 LDA 的文档生成

图中,每个圆圈表示一个随机变量,矩形表示变量的重复,同时参照其在生成过程中所扮演的角色进行标注。白色圆圈表示隐含变量,观测变量 $W_{d,n}$ 则用灰色的圆圈表示; D 表示文档的集合; K 表示设置的主题数目; a 表示每篇文档的主题分布的先验分布–Dirichlet 分布的超参数; η' 表示每个主题的词分布的先验分布–Dirichlet 分布的超参数; W 表示建模过程中可以观测的词语。具体的过程如下:

(1)从 Dirichlet 分布 a 中取样生成文档 d 的主题分布 θ_d 。

(2)从主题的多项式分布 θ_d 取样生成文档 d 第 n 个词的主题 $Z_{d,n}$ 。

(3)从 Dirichlet 分布 η' 中取样生成主题 $Z_{d,n}$ 对应的词语分布 β_k 。

(4)从词语的多项式分布 β_k 中采样最终生成词语 $W_{d,n}$ 。

用吉布斯采样法(Gibbs sampling)[7]对 LDA 模型的文档–主题分布和主题–词语分布进行推断,吉布斯采样的算法流程描述如下:

(1)初始化,对第 i 个词 W_i 随机分配某个主题。

(2)状态更新,对每个单词 W ,计算除 i 以外的其他全部词语的主题 $z-i$ ($-i$ 是 i 的补集)已知的情况下, W_i 属于每一个主题 j 的后验概率 $p(zj=j \mid z-I, w)$,将当前词语安排给概率值最高的主题。

(3)将第 2 步进行多次迭代,直到每个词语的主题收敛到稳定的状态。

1.3 时间因子

学术文献的一个重要属性是发表时间,发表时间越久,被引的数量越多,而发表时间越久,反而造成其热度下降,其时效性的特点不同于其他一些属性的文本,忽略时间容易造成主题挖掘不准确,即主题聚类的结果不正确。现有的主题分析模型[8-9]没有对学术文献的发表时间进行分析,而学术文献热点主题是具有时效性的,它随着时间的变化而变化,如果忽略这个特点,会导致主题分析的不准确性;每个学术文献都有自己的发表时间,如果学术文献的发表时间与当前时间的间隔越小,越能反映这一时间段内的学术热点主题,因此时间因素在考虑学术文献的主题上是不可忽视的因素[10]。针对这种缺陷,引入时间因子,根据德国心理学家艾宾浩斯提出的艾宾浩斯遗忘曲线来得到学术文献摘要的时间权重大小。将每个学术文献摘要的特征词根据发表时间权重分别相加,并按照权重和进行排序,然后用来训练时间窗口的大小,得出的时间窗口对学术文献主题分析的时间做出限定,发表时间位于在时间范围内的学术文献,对其摘要进行主题分析。

(1)构造学术文献的发表时间因子函数(如式 2),计算学术文献发表时间和当前时间的间隔,及其对学术文献主题的影响:

$$\mathrm{TimeWeigh} = \mathrm{e}^{-\frac{T_{\mathrm{now}}-T_{\mathrm{pub}}}{\mathrm{EWeigh}}} \tag{2}$$

其中,TimeWeigh 表示发表论文的时间和当前时间的时间差,以及时间差所反映的主题变化的权重; T_{now} 表示当前时间; T_{pub} 表示学术文献的发表时间; EWeigh 表示学术文献根据发表时间这一特点,得出的时间内主题的衰减因子,主题的衰减因子是由艾宾浩斯曲线拟合出的函数决定的。

(2)艾宾浩斯曲线是以一位心理学家的名字命名的。德国心理学家艾宾浩斯,通过研究人脑,发现人脑对于新事物的遗忘总是遵循着一种规律,这种规律可以由一种曲线所反映。在人们接触一种新鲜事物时,经过一阶段对于这种新鲜事物产生认识后,遗忘立刻开始,最初遗忘的速度很快,并且遗忘的数量很大,随着时间的变化,遗忘速度会变慢,遗忘的内容会减少,最终到达一定的程度,总结下来就是速度由快变慢,内容由多变少,这些都是德国心理学家艾宾浩斯的理念。这一过程的发现对于人类的记忆力研究有很大帮助,还能适用于多个领域[11]。文中将学术文献的特征词当作准备被新认知的事物,即是对应于人脑即将会产生记忆的材料,而计算机对应于人脑,会对这些特征词产生记忆,这个记忆的遗忘过程遵循艾宾浩斯遗忘曲线,对于特征词的遗忘情况进行记录,将结果拟合成函数,如式 3。

$$\mathrm{EWeigh} = 97.53\,(T_{\mathrm{pub}})^{-0.446} + 17.68 \tag{3}$$

(3)对学术文献摘要的发表时间进行分析,判断其是否在时间范围内,对于窗口范围内出现的学术文献摘要计算发表时间权重,依据式 3 将计算出来的发表时间权重进行求和运算,都是以特征词为单位而进

行的,得出学术文献摘要的某一个特征词的总的发表时间权重,如式 4:

$$\text{SumWeigh} = \sum_{T_{re}}^{T_{Fir}} \text{TimeWeigh}_i \quad (4)$$

其中, T_{re} 表示特征词离现在时间最近,出现的时间; T_{Fir} 表示特征词第一次在文档集中出现的时间; SumWeigh 表示各特征词的发表时间的权重和。

(4)学术文献的另一个属性是其拥有发表的作者,有如下情况,作者相同的学术文献,方向不同;作者相同的学术文献,方向相同;作者不同的学术文献,方向相同;作者相同的学术文献,方向不同。综合上述因素考虑,将学术文献的摘要以作者为区分变量进行分类,并且建立目标文档集,建立文档集后,对文档集内的每一篇学术文献摘要进行预处理,处理主要有分词和去除停用词,并且统计每一篇学术文献摘要的发表时间,以便计算发表时间的权重。这样的学术文献摘要内容才能更加适用于特征提取算法,将分词和去除停用词的学术论文摘要内容使用 TTF-IDF 进行特征提取,提取出可以代表学术文献摘要内容的特征词,对内容的数量进行简化,同时对学术文献摘要的发表时间进行转化,将其表示成二元组的形式,使其序列化。二元组<word,time>中,word 表示学术文献摘要中的某个特征词,time 表示该特征词所在的学术文献的发表时间。

设置学术文献的发表时间窗口分为以下几步:

(1)根据式 3,可得出每一篇学术文献摘要的发表时间权重大小,对应于一个一个的点,这些点的斜率值就是学术文学摘要的发表时间权重。在三角符号 93 天处,发现斜率的变化小于 0.02,此时对于特征词的记忆程度的遗忘是一个很重要的时间点。人脑对于学术文献特征词的遗忘趋于平稳,遗忘的速度和量将不会发生大的改变,因此将学术文献摘要所对应的时间窗口初始化为 93 天,如图 2 所示。

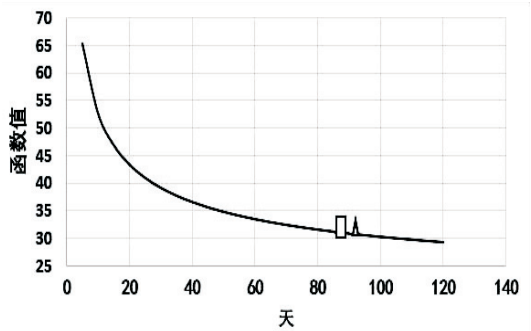

图 2　时间窗口

(2)将学术文献的发表时间数据以月为单位进行转化,转化后进行抽取,抽出 40%,将这 40% 的学术文献摘要作为训练集,训练叙述学术文献发表时间所对应的时间窗口大小。

(3)对训练集进行计算,得出训练集的每一篇学术文献摘要所对应的每一个特征词的发表时间权重,并对这些特征词的总的发表时间权重进行计算,即 SumWeigh。将各个特征词按得出的 SumWeigh 值进行降序排列,取 SumWeigh 值大的前 100 个特征词进行记录,记为 T_1,并且定义一个变量 j,用来对学术文献摘要发表时间所对应的时间窗口进行操作,初始化 $j=0$。

(4)学术文献的发表时间所对应的时间窗口的大小减少 $\Delta t(\Delta t = 2^j)$,对训练集进行计算,将各个特征词按得出的 SumWeigh 值进行降序排列,取值为前 100 个特征词记为 T_2, j ++。

(5)按照 T_1 和 T_2 中的相同特征词数目计算 T_1 和 T_2 的匹配度,记为 M。

(6)若 M <0.8(特征词的相同量小于 80%),认为收敛,确定时间窗口大小为 93- Δt,否则,将 T_2 集合覆盖为空(NULL),返回步骤 4。

2　基于学术文献的 TF-LDA 主题模型

学术文献文本具有时效性的特点,在分析其主题时需要考虑各文本的发表时间,而 LDA 模型的本质是显示出主题的概率,其是一种主题概率模型,忽略词序、语法等,认为每个词与每个词之间是独立的,没有联系,可以独立出现,在任意位置选择一个词都不会受到前面选择的影响。词知识跟该词所处的主题有关,在建模的过程中,以词频作为基础[12-13],词频高的词就有优势,对主题进行选择时,会偏向高概率词。而在学术文献的摘要中并不是出现次数多的词就一定是学术文献摘要的特征词,能代表学术文献的摘要内容,显然这种主题选择方式对于学术文献的摘要并不适合,不符合学术文献摘要的主题分布,并且 LDA 模型提取主题时,没有将学术文献摘要的重要因素—发表时间考虑在内,使效果不佳,挖掘出来的学术文献摘要的主题不符合学术文献摘要的内容。

综上所述,在进行学术文献摘要的主题分析时应加入学术文学摘要的发表时间,并且对 LDA 建模过程中的采样策略进行改进,然后进行学术文献摘要的主题挖掘。具体如下:使用 TF-IDF 提取特征词,进行初步采样,形成一个主题引导特征词词库,对主题引导特征词词库进行计算,得到特征词的权重,使用主题引导特征词词库进行引导,从而促使主题的提取更加准确,达到增加主题引导特征词词库对主题建模产生作用的状态。并提出发表时间因子,将每个学术文献的发表时间作为其时间标签,在特征词分配给主题的过程中,利用时间因子产生的时间窗口进行时间限制,优化主题的选取,增加发表时间影响权重的大小,距离当前时间越近的特征词,所对应的时间权重就应该越大,从而符合学术文献摘要的发表时间特点。改进后的总体步骤如下:

(1)输入文档集合,进行分词和去除停用词等预处理;

(2)根据 TF-IDF 提取特征词;

(3)初步采样;

(4)特征词标注;

(5)构建主题引导特征词词库;

(6)综合步骤 3、5,计算特征词引导权重;

(7)利用艾宾浩斯遗忘曲线进行时间权重的计算;

(8)综合步骤 6、7,计算总的影响权重;.

(9)利用吉布斯采样算法对分词后的文本数据进行迭代采样;

(10)迭代完成,输出主题模型的结果。

将学术文献的发表时间融合到 LDA 模型中,对 TTF-LDA 模型中的词条进行表示,对于学术文献摘要的特征词的发表时间,以及词条的表示形式如图 3 所示。

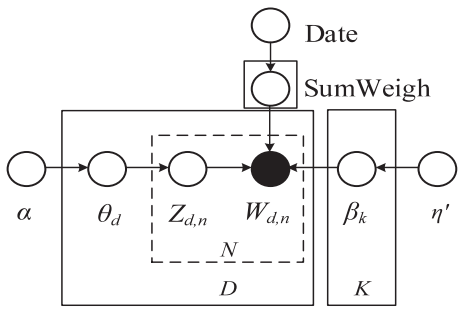

图 3 融合发表时间的 TTF-LDA 模型

在模型中融合发表时间因子后,学术文献文本中各特征词的概率分布可以展现出来,如式 5 所示:

$$p(W_m \mid Z_m) = \sum_{<W_m,t> \in W} p(t)p(W_m \mid t, Z_m) \tag{5}$$

用吉布斯采样进行推理,推理结果的特征词和主题服从的分布如式 6 所示:

$$p(\vec{W}\mid_m, \vec{Z}\mid_m \mid \vec{\alpha}, \vec{\beta}) = p(\vec{W}_m \mid \vec{Z}_m, \vec{\beta})p(\vec{Z} \mid \vec{\alpha}) \tag{6}$$

其中,$p(\vec{W}\mid_m \vec{Z}\mid_m \vec{\beta})$ 表示在一种先验分布情况下,参数为 $\vec{\beta}$ 以及特征词主题为 $\vec{Z}\mid_m$,对特征词进行的采样;$p(\vec{Z}\mid\mid \vec{\alpha}\mid)$ 表示在另一种先验分布情况下,参数为 $\vec{\alpha}$ 以及特征词主题为 $\vec{Z}\mid_m$,对学术文献摘要主题进行的采样。结合这两个因子进行计算,可以得到学术文献摘要的后验估计。

特征词的采样独立进行,和主题的采样没有联系,同样主题的采样也是独立进行的,和特征词的采样也没有联系,并对这两个过程进行计算。求解第一个因子,在参数为 $\vec{\beta}$ 的先验分布和主题为 $\vec{Z}\mid_m$ 的条件下,采样得到的特征词分布 $\vec{\ell}$ 和特征词中服从主题 Z 的多项分布如式 7 所示:

$$p(\vec{W}\mid_m \mid \vec{Z}\mid_m, \vec{\ell}) = \prod_{n=1}^{N} p(\vec{W}\mid_n \mid \vec{Z}\mid_n) = \prod_{n=1}^{N}\prod_{t=1}^{T} \ell_{z,t}^{n_z^{(t)}} \tag{7}$$

其中,$n_z^{(t)}$ 表示特征词 t 在时间因子的作用下分给主题 Z 的过程中的权重大小。结合时间权重,将其与时间权重的乘积作为新的权重,得出如下结果:

$$n_z^{(t)} = n_z^{(t)} \cdot \text{TimeWeigh} \tag{8}$$

则第一项因子的最后结果为:

$$p(\vec{W}_m \mid \vec{Z}\mid_m, \vec{\ell}) = \prod_{z=1}^{N} \frac{\Delta(\vec{n}_z + \vec{\beta})}{\Delta(\vec{\beta})}, \vec{n}_z = \{n_z^{(t)}\}_{t=1}^{T} \tag{9}$$

第二项因子 $p(\vec{Z} \mid \vec{\alpha})$ 表示在特征词主题为 $\vec{Z}_m$ 且先验分布参数为 $\vec{\alpha}$ 时对主题进行的采样,主题采样时不需要加入时间因子,其推导结果为:

$$p(\vec{Z} \mid \vec{\alpha}) = \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})}, \vec{n}_m = \{n_m^{(z)}\}_{z=1}^{Z} \tag{10}$$

其中,$n_m^{(z)}$ 表示主题为 z 的情况下,在文章 m 中出现的次数。将两个因子的计算结果进行乘法运算,得

到如式 11 的学术文献摘要的主题模型的联合分布:

$$p(\vec{W}_m, \vec{Z}_m \mid \vec{\alpha}\vec{\beta}) = \prod_{Z=1}^{N} \frac{\Delta(\vec{n}_z + \vec{\beta})}{\Delta(\vec{\alpha})} \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})} \tag{11}$$

学术文献摘要的超参数为 α_k 和 β_t ,后期期望如式 12 和式 13 所示:

$$\ell_{z,t} = \frac{n_z^{(t)} + \beta_t}{\sum_{t=1}^{T} n_z^{(t)} + \beta_t} \tag{12}$$

$$\theta_{m,z} = \frac{n_m^{(z)} + a_k}{\sum_{z=1}^{Z} n_m^{(z)} + a_k} \tag{13}$$

其中, $\ell_{z,t}$ 表示主题为 z 的情况下,特征词 t 处于主题 z 内的概率; $\theta_{m,z}$ 表示对于文档 m,如果其主题为 z,而主题 z 的概率。

依靠联合分布以及后验期望,对隐含变量也就是需要的学术文献摘要的主题 Z ,可以在考虑发表时间因素下,挖掘出学术文献摘要中隐藏的主题,得到主题的分布。

3　实　验

3.1　实验数据及预处理

实验数据采用爬虫爬取的知网上的论文摘要共 46 312 条,在数据预处理阶段首先对摘要的标点符号进行去除,将纯文本数据使用 python 的 jieba 库进行分词,并去除停用词,将分词和去除停用词后的文本数据整合成文档。

3.2　实验结果与分析

文中提出的 TTF-LDA 主题模型的参数设置为主题数 $K=20$,超参数 $\alpha=1$, $\beta=0.02$, $\delta=0.02$,吉布斯采样的迭代次数一般设置为 2 000。在初始时间窗为 93 天的情况下计算学术文献的发表时间权重,图 2 中正方形处得出时间窗口大小为 86 天。TTF-LDA 模型的 6 个主题结果如图 4 所示。

Topic1		Topic2		Topic3	
人工智能	0.0879253	机器人	0.0756232	计算机视觉	0.0682348
技术	0.0623782	智能	0.0645232	应用	0.0563481
领域	0.0583458	传统	0.0542365	深度学习	0.05512376
应用	0.0532412	系统	0.0423562	技术	0.0478958
发展	0.0478236	自动化	0.0412321	识别	0.0356738
理论	0.3563475	设计	0.0324578	目标	0.0348623
Topic4		**Topic5**		**Topic6**	
机器视觉	0.0675987	深度学习	0.0736231	机器学习	0.0671252
计算机	0.06245826	神经网络	0.0714567	医学	0.0621243
人工智能	0.0523489	算法	0.0685423	算法	0.0672361
发展	0.04689563	人工	0.0526112	实现	0.0512342
处理	0.04569321	模型	0.0451236	分类	0.0456231
工作	0.03214567	综述	0.0321523	教育	0.0421237

图 4　部分主题结果

6 个主题分别是有关人工智能、机器人、计算机视觉、机器视觉、深度学习和机器学习,在 TTF-LDA 的权重值中人工智能这一特征词的权重最高,是在 2017-10-1 至 2017-12-31 期间发生的最热门的主题,也符合趋势。目前人工智能的发展最为普遍,人工智能类的论文也最多,机器人相关的论文数量也很多,而深度学习和机器学习为人工智能领域下的两大热点话题,概率也高于其他话题,主要是相同的特征词在时间权重的影响下权重更高,反映出文中模型能准确挖掘出相关主题。在主题模型中,主题与主题之间的相似性越低则效果越好,图 5 为 TTF-LDA 和 LDA 模型的主题之间的相似度的对比情况。结果表明,在文档集增加的情况下,主题之间的相似度在降低,但 TTF-LDA 的主题之间的差异大,效果优于 LDA,主要是对主题引导词的加权,提高主题引导词在文档中的重要性,特征词引导主题的贡献也越大,使得结果更加符合文档集自身的分布特点,主题提取更准确。

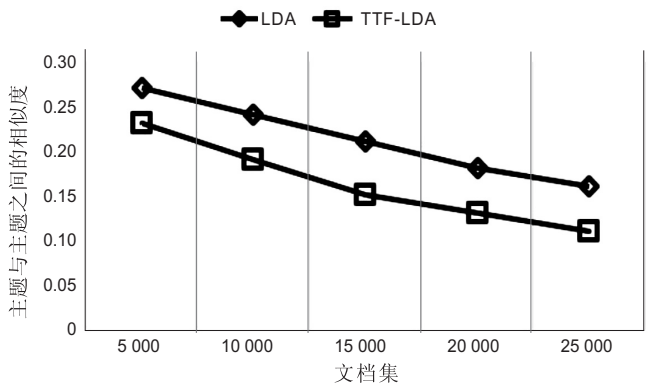

图 5　主题之间的相似度

对 TTF-LDA 模型、LDA 模型和 WMC-LDA 模型进行对比分析,使用评价指标混乱程度,用 Perplexity 值来代表主题分析后的情况,对主题分析后的情况进行混乱程度对比和分析。混乱程度是主题模型研究中常用的对比指标[14]。在混乱程度的理念里,如果 Perplexity 越大,则表示这个主题模型的混乱程度越混乱,效果越差,与之相反,如果 Perplexity 值越小,则表示这个主题模型的混乱程度越小,即主题很清晰,效果越好。Perplexity 的定义如式 14 所示。

$$\text{Perplexity}(W) = \exp\left\{-\frac{\sum_{m} \text{Inp}(W_m)}{\sum m \cdot N_m}\right\} \tag{14}$$

其中,W 为测试集,由学术论文摘要组成;W_m 为测试集中抽取到的特征词,对应于学术文献摘要内容由 TF-IDF 提取出的,能表示学术文献摘要的特征词;N_m 为特征词的总数,统计所有的特征词总数得来。

TTF-LDA、LDA 和 WMC-LDA[15] 的 Perplexity 与迭代次数的关系如图 6 所示,实验的条件都设置一样,其中纵坐标为 Perplexity/100。

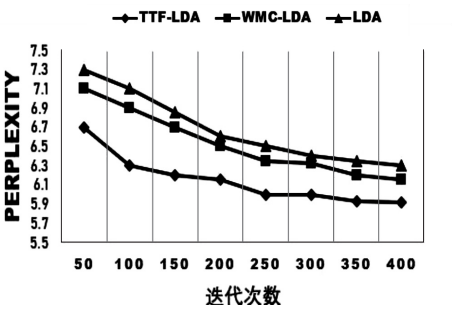

图 6 Perplexity 值

由图 6 可以看出,在其他情况都相同的条件下,随着迭代次数的增加,三种模型的 Perplexity 值都一直下降,而 TTF-LDA 模型的 Perplexity 值一直都最小,表明其运算速度更快、效率更高,内部的主题提取情况更加明确,证明提出的采样的策略和增加的学术文献的发表时间权重是有效的。

4 结束语

文中提出一种学术文献摘要的主题分析方法,针对现有的主题分析中的主题模型未考虑论文发表时间的缺点,提出将学术文献的发表时间适用于人脑的记忆遗忘规律,使遗忘曲线计算出学术文献特征词的遗忘曲线,设置学术论文摘要的发表时间对应的时间窗口,对主题的时间范围进行缩短,并利用特征词处理后得到的主题引导特征词词库,共同引导主题分布。通过实验证明了该方法的可行性,能准确地挖掘出当前学术文献摘要的主题。

参考文献:

[1] 胡 侠,林 晔,王 灿,等. 自动文本摘要技术综述[J]. 情报杂志,2010,29(8):144-147.

[2] 刘天祎,步 一,赵丹群,等. 自动引文摘要研究述评[J]. 现代图书情报技术,2016,32(5):1-8.

[3] 施聪莺,徐朝军,杨晓江. TF-IDF 算法研究综述[J]. 计算机应用,2009(S1):167-170.

[4] 胡 亮,夏 磊,李 伟. 基于改进 TF-IDF 算法的关键词抽取系统[J]. 厦门理工学院学报,2017,5(3):12-16.

[5] 唐晓波,王洪艳. 基于潜在狄利克雷分配模型的微博主题演化分析[J]. 情报学报,2013,32(3):281-287.

[6] 余维军,刘子平,杨卫芳. 基于改进 LDA 主题模型的产品特征抽取[J]. 计算机与现代化,2016(11):1-6.

[7] BORNN L,CHEN Y,FREITAS N D,et al. Herded Gibbs sampling[J]. Journal of Machine Learning Research,2013,17:26-32.

[8] 王立人,余正涛,王炎冰,等. 基于有指导 LDA 用户兴趣模型的微博主题挖掘[J]. 山东大学学报:理学版,2015,50(9):36-41.

[9] 谢 昊,江 红. 一种面向微博主题挖掘的改进 LDA 模型[J]. 华东师范大学学报:自然科学版,2013,20(6):93-101.

[10] 韩忠明,陈 妮,乐嘉锦,等. 面向热点话题时间序列的有效聚类算法研究[J]. 计算机学报,2012,35(11):2337-2347.

[11] 于 洪,李转运. 基于遗忘曲线的协同过滤推荐算法[J]. 南京大学学报:自然科学版,2010,46(5):520-527.

[12] TANG Jie,WANG Bo,YANG Yang,et al. Patentminer:topic-driven patent analysis and mining[C]//Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. Beijing:ACM,2012:1366-1374.

[13] VORONTSOV K,POTAPENKO A. Additive regularization of topic models machine learning[J]. Machine Learning,2015,101(1-3):303-323.

[14] 张晨逸,孙建伶,丁秩群. 基于 MB-LDA 模型的微博主题挖掘[J]. 计算机研究与发展,2011,48(10):1795-1802.

[15] 李 鹏,于 岩,李英乐,等. 基于权重微博链的改进 LDA 微博主题模型[J]. 计算机应用研究,2016,33(7):2018-2021.