

基于 LDA 的社科文献主题建模方法

李昌亚, 刘方方

(上海大学 计算机工程与科学学院, 上海 200444)

摘要: 随着互联网的发展, 文本分类和主题提取的应用越来越广泛, 而主题模型在文本主题提取中起着很大的作用。LDA (latent Dirichlet allocation) 模型是一种应用非常广泛且很成熟的主题模型, 也是一个概率生成模型, 可以很好地解决多词一义和一词多义的问题。但是当利用 LDA 模型对社科文献领域类的文档集进行主题建模时, 由于该建模方法忽略了文档集自身的主题特点, 提取的主题分布是偏向文档中高频词汇, 所以造成最后提取的主题偏离文档的本质意义上的主题, 结果不够准确。针对 LDA 模型对文档进行主题建模的过程, 结合社科文献领域的文档特点, 对主题建模的过程进行相应的改进, 提出一种新的主题建模方法, 从而使最终提取的主题更加准确, 更符合文档集本身的主题特点。

关键词: 主题模型; LDA; 社科文献; Gibbs 抽样

中图分类号: TP31

文献标识码: A

文章编号: 1673-629X(2018)02-0182-06

doi: 10.3969/j.issn.1673-629X.2018.02.039

A Topic Modeling Method for Social Science Literature Based on LDA

LI Chang-ya, LIU Fang-fang

(School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China)

Abstract: With the development of the Internet, the application of text classification and topic extraction is becoming more and more widely, and topic model plays a critical role in topic extraction of the text. LDA (latent Dirichlet allocation) as an extensive and mature topic model is also a probability generation model, which can solve the problem of synonym and polysemy. But when LDA model is used to model the document collection in the domain of social science literature, because of its ignorance of the topic characteristics of document collection itself, the topic distribution extracted by the modeling method is to trend the high frequency words, which makes the extracted topic deviated from the document topic in nature and the results inaccurate. In this paper, aiming at the topic modeling of document with LDA model and combined with the characteristics of the document in the field of social literature, we present a new topic modeling method to improve accordingly the process of modeling, so that the topic of the final extraction is more accurate and more consistent with the topic characteristics of the document collection itself.

Key words: topic model; LDA; social science literature; Gibbs sampling

0 引言

社会科学文献, 简称社科文献, 是指记载有关社会科学知识、信息的一切物质载体。在现代信息社会中, 社科文献数量庞大, 内容广泛, 种类繁多, 载体多样, 发展速度快^[1], 尤其是社科文献领域下的专题文献主题交叉比较明显。随着社会的发展, 文献中会不断出现很多新词, 很多词语会具有特定含义, 如“四化并举”、“黄金水道”、“成本化解”等。这种情况仅仅通过人工分类, 或者单纯地依靠机器自动进行主题提取、分类, 不能很好地提取出文献的主题和类别。

对于处理自然语言中的问题, 主题模型是一种很常见和成熟的技术。主题模型的起源是隐性语义索引 (latent semantic analysis, LSA)^[2], 严格意义上讲, 隐性语义索引并不是真正的主题模型, 但是其基本思想促进了主题模型的发展。概率隐性语义索引 (probabilistic latent semantic analysis, pLSA)^[3] 就是由 LSA 发展而来的一个基于概率模型的主题模型。Blei 等以 pLSA 为基础提出的 LDA (latent Dirichlet allocation)^[4] 是一个完全的概率生成模型。近年来, 业内出现的许多概率模型都是以 LDA 为基础, 结合不同的业务进行改

收稿日期: 2017-03-01

修回日期: 2017-07-06

网络出版时间: 2017-11-15

基金项目: 上海市科委自然科学基金 (12ZR1411000)

作者简介: 李昌亚 (1991-), 男, 研究生, 研究方向为信息检索、自然语言处理; 刘方方, 博士, 副教授, 研究方向为知识表示与推理、Web 服务匹配。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20171115.1128.018.html>

进的,但是这些算法都不太适合对社科文献领域类的文档集进行主题提取。

因此,为了能够高效、准确地提取社科文献领域类中文档集的主题,文中提出一种以 LDA 模型为基础,结合社科文献的特点,改进主题建模过程的主题建模方法。

1 相关工作

LDA 主题模型,本质思想是以概率为基础对文本进行主题建模。它独立于文本自身特点,所以对于不同领域的业务需求,如果直接应用 LDA 算法进行主题建模,结果都是不够精准的。因此,提出了很多结合不同的业务需求,对 LDA 进行相应改进的算法。首先是将 LDA 模型运用到短文本领域,如微博、用户评论等,它们对 LDA 模型本身没有过多的改进,而是偏向对 LDA 模型的应用。张志飞等^[5]利用 LDA 建模后的主题特点对短文本进行分类;高明等^[6]将 LDA 主题模型应用在对微博的个性推荐方面。将 LDA 模型应用到博客、帖子和话题追踪和预测等方面^[7-8]都有很好的价值,但是对于该论文背景中篇幅较长、量比较大的社科文献,这些方法明显不太适合。

其次,主要是在 LDA 模型的基础上引入新的参数或者约束条件。何锦群^[9]提出对文档集的所有隐藏主题进行分类,即主题层之上再引入一层表示主题类别,该算法适合文档集主题分布比较广泛的情形,但是对于主题交叉和特征词比较明显的文档集效果不太明显。SA-LDA^[10]算法和 SRC-LDA^[11]中利用句法分析构建语料库的约束条件,从而引导主题建模,但该算法使用于短文本,不太适合具有篇幅较长的文档。

另外,还有就是从 LDA 概率模型的基本原理出发,认为概率模型主要受高频词的影响,会使得建模后的主题不够明确。胡勇军等^[12]利用 LDA 高频词作为短文本分类的空间模型的特征向量。张小平^[13]在建模过程中选择降低高频词的权重,这个改进虽然可以降低常用高频词汇对建模的影响,但是对于那些主题词比较明显、具有很多新词新义的文档集依然存在不足。虽然从整体上降低了常用高频词的干扰,但是不能提高特征词和新词在主题建模过程中的重要性。

因此,文中提出一种根据文档集自身的主题特征进行特征词标注,然后在 LDA 建模过程中增加主题特征词权重的建模方法,从而使得建模结果的主题分布更加准确,更加符合文档集自身的特点。

2 LDA 模型

2.1 模型描述

LDA 是一种非监督学习技术,可以用来识别大规

模文档集(document collection)或语料库(corpus)中隐藏的主题信息^[14-15]。LDA 模型从实际情况出发,一个文档由多个隐含主题随机组成,而每个主题又可以由文档中的多个词语进行表示(如图1所示)。因此,可将一篇文档表示为隐含主题的概率分布(doc-topic),而每一个隐含主题又可以看作词语的概率分布(topic-word)。这种思想有利于大规模文档集处理中的空间降维,即把文档映射到 topic 层面上。LDA 在建立两个分布时,采用了词袋(bag of words)^[16]的方法,这种方法忽略了每一篇文档中句子的语法、次序,以及词之间的关系,文档中每个词语的出现都是相互独立的。

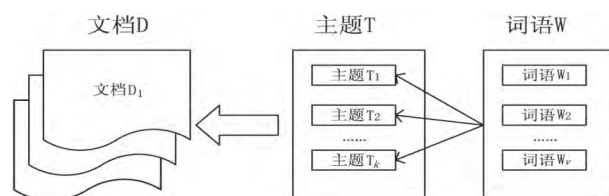


图1 文档-主题-词语关系

LDA 建模的基本层次结构是文档-主题-词语。简单理解为:每篇文档都是由若干个词语组成的,同时每篇文档都有潜在的几个相关的主题,而组成这篇文档所有的有用词语可以看成是这几个主题词对应的一部分相关的词语。因此,认为文档与主题之间的概率分布(doc-topic)是多项分布 $Z \sim \text{Multinomial}(\theta_m)$, 主题与词语之间的概率分布(topic-word)也是多项分布 $W \sim \text{Multinomial}(\psi_k)$ 。LDA 引入了 Dirichlet 分布作为多项分布的先验分布进行求解,LDA 模型的结构如图2所示。

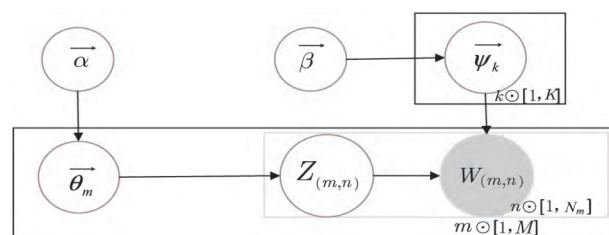


图2 LDA 模型结构

图中, M 表示文档集中的文档数; K 表示设置的主题数; V 表示文档集中出现的所有词且不重复的词表; θ 表示一个 $M \times K$ 的矩阵, θ_m 表示第 m 篇文档的主题分布; ψ 表示一个 $K \times V$ 的矩阵, ψ_k 表示编号为 k 的主题之上的词分布; α 表示每篇文档的主题分布的先验分布-Dirichlet 分布的超参数,其中 $\theta_m \sim \text{Dir}(\vec{\alpha})$; β 表示每个主题的词分布的先验分布-Dirichlet 分布的超参数,其中 $\psi_k \sim \text{Dir}(\vec{\beta})$; W 表示建模过程中可以观测的词语。

利用 LDA 模型^[17]对文档集主题模型进行生成可以理解为对整个文档集中的词语进行生成,其具体过

程如下:

(1) 建模过程中, 首先给定 α 和 β 的值, 及主题 K 的取值。对于 α 和 β 通常根据经验取值。

(2) 确定文档的主题。首先, 根据 α 并结合 $\vec{\theta}_m \sim \text{Dir}(\vec{\alpha})$ 确定文档的主题分布 $\vec{\theta}_m$, 对于第 i 篇文档, 主题分布为 $\theta_i = \{\theta_{i1}, \theta_{i2}, \dots, \theta_{iK}\}$; 然后, 由主题分布 θ_i 确定第 i 篇文档中第 j 个词的主题 $z_{ij} = k, k \in [1, K]$ 。

(3) 确定文档中的词语。首先, 在给定主题 z_{ij} 的情况下, 根据 β 值结合 $\vec{\psi}_k \sim \text{Dir}(\vec{\beta})$ 确定对应主题的词分布 $\vec{\psi}_k$, 对于第 i 篇文档中的第 $z_{ij} = k$ 个主题的词分布为 $\psi_k = \{\psi_{k1}, \psi_{k2}, \dots, \psi_{kv}\}$; 然后, 由词分布 ψ_k 确定具体的词语, 即可以得到观测值 w_{ij} 。

(4) 重复步骤 2 和步骤 3, 直到完成对一篇文档的所有词语的生成, 再到所有文档的生成。

由 LDA 模型的建立过程可知, 对于图 2 中参数的理解, 可以得到所有变量的联合分布公式:

$$P(W_m, Z_m, \theta_m, \psi_k | \alpha, \beta) = \prod_{n=1}^{N_m} p(W_{m,n} | \psi_{z_{m,n}}) p(Z_{m,n} | \theta_m) p(\theta_m | \alpha) p(\psi_k | \beta) \quad (1)$$

利用上述 LDA 模型对社科文献领域的文档集建模后的部分主题结果如表 1 所示。结果表明, 同一个主题下的所有主题词明显是同一类别的词语, 但是部分主题中仍然存在一些相关度很低的词语, 而且对于一些文档集中出现次数较少且很重要的词语-主题特征词, 明显是不会出现在主题词中或者位序很靠后。

表 1 LDA 模型建模后的主题类别

主题 1	主题 2	主题 3	主题 4	主题 5	主题 6
生态	产业	旅游	金融	人才	区域
环境	服务业	商业	贷款	创新	长三角
资源	制造业	创意	资金	科技	中心
循环	工业	旅游业	银行	信息	江苏
能源	行业	品牌	融资	研究	一体化
利用	产品	杭州	中小企业	培训	合作
持续	集群	休闲	风险	知识	南京
江西	物流	文化产业	投资	创业	资源
污染	优势	资源	金融机构	高校	产业
生产	生产	保护	业务	技术创新	优势
环保	竞争力	特色	保险	机构	协调
排放	产业结构	产品	机构	开发	上海
节能	集聚	开发	经营	能力	长江三角洲
绿色	轻工业	历史	信用	建立	宁波
保护	升级	世博会	资本	知识产权	战略

2.2 模型实现方法

LDA 模型的超参数估计通常采用 EM 算法和

Gibbs Sampling 算法, 其中 Gibbs Sampling 算法通过迭代采样来逼近真实的概率分布^[18], 实现相对简单, 而且应用较广泛。

在实际应用过程中, 从文档集的输入到最终主题模型结果的输出, 对文档集预处理、分词之后直接应用 Gibbs 采样对文档集进行 LDA 模型的实现处理, 其步骤如下:

Step1: 输入文档集进行预处理、分词;

Step2: 利用 Gibbs Sampling 算法对分词后的文本进行迭代采样;

Step3: 迭代完成, 输出主题模型结果。

整个模型实现过程并没有考虑到文档集自身的特点, 而是对预处理、分词之后的文档集直接利用 Gibbs Sampling 算法进行实现。这种做法会造成主题分布偏向文档中那些常用的高频词, 忽略了文档中那些主题特征明显的词语在建模过程中的影响。

根据式(1)可知, 由于 \vec{W} 是观测到的已知数据, 只有 \vec{Z} 是隐含的变量, 所以只需要利用 Gibbs Sampling 采样 $p(\vec{Z} | \vec{W})$ 。

$$P(z_i = k | \vec{Z}_{-i}, \vec{W}) \propto \frac{n_{m, \cdot i}^{(k)} + \alpha_k}{\sum_{k=1}^K (n_{m, \cdot i}^{(k)} + \alpha_k)} \cdot \frac{n_{k, \cdot i}^{(i)} + \beta_i}{\sum_{t=1}^V (n_{k, \cdot i}^{(t)} + \beta_t)} \quad (2)$$

其中, $n_{m, \cdot i}^{(k)}$ 表示第 m 篇文档主题词 k 的个数去掉一个后的数目; $n_{k, \cdot i}^{(i)}$ 表示第 k 个主题下词语 i 的个数去掉一个后的数目。

利用 Gibbs 采样后得到两个分布: 文档与主题的分分布 (doc-topic) $\vec{\theta}_m$ 和主题与词语的分分布 (topic-word) $\vec{\psi}_k$:

$$\theta_{m,k} = \frac{n_{m,k} + \alpha}{\sum_{i=1}^K n_{m,i} + K\alpha} \quad (3)$$

$$\psi_{k,w} = \frac{n_{k,w} + \beta}{\sum_{i=1}^V n_{k,i} + V\beta} \quad (4)$$

在实际处理过程中, 超参数 α 和 β 是作为常量处理的。式(3)表示文档 m 下的主题 k 的分布情况, 其中 $n_{m,k}$ 表示文档 m 下主题 k 出现的次数, 分母表示文档 m 中所有主题出现次数的总和。式(4)表示主题 k 下的词 w 的分布情况, 其中 $n_{k,w}$ 表示主题 k 下的词 w 出现的次数, 分母表示主题 k 中所有词语出现次数的总和。

根据 Gibbs Sampling 的公式可以得出, $n_{m,k}$ 和 $n_{k,w}$ 是对所有的词或者主题通过词自身出现的次数或者主

题被记录次数的统计,而并没有考虑词本身对该文档的重要性。由于高频词在主题中和文档中占有的比例都较大,导致主题的分布偏向高频词的主题倾斜。对于那些主题特征比较明显的词语,出现的次数比较少,就会在建模过程中低作用化,甚至没有作用。

如果在利用 Gibbs 采样过程中,对于采样那些主题特征比较明显的词语时,不仅考虑统计次数,而且考虑它们在文档中的权重值,那么就能增加这些词语在建模过程中的重要性。如表1中的主题3的“文化产业”和“世博会”、主题4的“金融机构”、主题5的“知识产权”、主题6中的“长江三角洲”等词语,如果增加这些词语的权重值,那么在模型结果中,它们对应的主题词位序就会上升,主题词间相关度也会增加。

3 改进的 LDA 模型建模方法

由于 LDA 模型是一种概率模型,建模过程中是以词频作为基础,所以对主题的采样结果会偏向高频词。这种建模方法是不符合社科文献主题分布特点的。为了能够对社科类文献提取更加准确的主题模型,提出一种结合文档集自身隐含的主题特征,改进 LDA 建模过程中采样策略的方法,然后应用该方法对文档集进行主题的提取。

3.1 主题建模过程

主要介绍的是对建模过程中的改进方法,不同于改进前的 LDA 建模过程。改进后的建模过程中将考虑文档集自身的主题特点,即在文档集预处理之后,先进行初步采样,根据文档集的特征词标记文档中的词语信息,形成一个主题引导词库,然后再利用主题引导词库计算特征词的权重,从而达到增加主题引导词对主题建模时的影响。与之前的三步实现过程比较,增加了相应的四个处理步骤,改进后的总体实现步骤如下:

Step1: 输入文档集进行预处理、分词;

Step2: 初步采样;

Step3: 特征词标注;

Step4: 获取主题引导词库;

Step5: 结合步骤2、4,计算引导词权重;

Step6: 利用 Gibbs Sampling 算法对分词后的文本进行迭代采样;

Step7: 迭代完成,输出主题模型结果。

在利用 Gibbs 采样之前对预处理、分词后的文档集进行初步采样、特征词标注、计算引导词权重三个过程。通过这三个过程可以提取出文档集中主题特征比较明显的词语信息,然后利用这些信息再进行 Gibbs 采样。

初步采样: 记录文档集中文档数量和每篇文档中词语的数量。

特征词标记: 标记每一个主题特征词在每一篇文档中出现的次数。该过程的结果形成一个主题引导词库,词库中的词都是文档中出现频率较小的,而且是文档主题导向的关键词。

计算引导词权重: 根据初步采样的信息和特征词标记的结果,计算主题特征词中每一个词在每篇文档中对应的权重值。

由于引导词库中不同的词语出现的频率不一样,而且对于不同文档的重要程度也不一样,故借鉴 TF-IDF^[19] 的思想和一个可变参数 δ 对引导词进行权值计算。

$$G_{m,t} = \frac{n_{m,t}}{\sum_{t=1}^T n_{m,t}} + \delta_t \quad (5)$$

其中, $G_{m,t}$ 表示第 t 个特征词在文档 m 中的权重; $n_{m,t}$ 表示第 t 个特征词在文档 m 中出现的次数; 分母表示文档 m 中的总词语数; $\vec{\delta} = \{\delta_1, \delta_2, \dots, \delta_t\}$ 是一个向量, δ_t 表示第 t 个特征词的一个引导参数,根据经验一般取值为 $0 \sim 0.1$ 。

计算好主题引导词对应的权重之后,在 Gibbs Sampling 过程中,在求主题和词的分布时加上该词对应的权重值即可。

由于该方法中考虑了文档集自身的特点,增加了主题特征词在采样过程中的重要性,所以会增加主题特征词在文档主题分布中的影响,最终使建模结果更加准确,文档集的主题分布更加精准。

3.2 主题建模方法的实现步骤

改进后的建模方法较改进前的建模过程在 Gibbs 采样前增加了一些与计算主题特征词相关的处理过程。因此,改进后算法的基本步骤如下:

步骤1: 输入分词后的每一篇文档,一行为一篇文档,同时输入 α 、 β 、 δ 、主题数 K ,以及迭代次数。

步骤2: 扫描每一篇文档的每一个词语和特征词库进行比较,并记录每篇文档词的个数 d_m ,如果不是特征词,循环此步骤,直至所有的文档都被扫描完。

步骤3: 标记特征词,同时在主题引导词库中记录特征词信息 $w_t^{(m)} + 1$,表示主题引导词库中的第 t 个词语是文档 m 中的,并且其频率增加一次。

步骤4: 初次扫描完所有的文档集后,根据主题引导词库中的信息 \vec{w} 和 \vec{d} 计算主题引导词库中的每个词语对应的权重。

步骤5: Gibbs 采样,采样过程中统计每篇文档中每个词语的频率时,要和特征词进行比对,如果是,需要在主题特征词中查找对应的权重,并且加上该词对应的权值。

步骤 6: 循环执行步骤 5, 直至结果收敛, 即迭代次数完成。

步骤 7: 输出主题建模结果。

变量 $\vec{d} = \{d_1, d_2, \dots, d_m\}$ 记录整个文档集中对应每篇文档中词总数的向量; $\vec{w}_t = \{w_t^{(1)}, w_t^{(2)}, \dots, w_t^{(m)}\}$, $w_t^{(m)}$ 表示第 t 个特征词在文档 m 中出现的次数。所以该算法在初步采样后就可以计算出所有的主题引导词对应的权重值, 故式 (5) 可以写成:

$$G_{m,t} = \frac{n_{m,t}}{\sum_{e=1}^V n_{m,e}} + \delta_t = \frac{d_m}{w_t} + \delta_t \quad (6)$$

因此在利用 Gibbs 采样时, 计算 (doc-topic) θ_m 和 (topic-word) ψ_k 遇到主题特征词时, 需加上特征词对应的权重。故式 (3) 和式 (4) 可以分别写成:

$$\theta_{m,k} = \frac{n_{m,k} + \sum_{t=1}^T G_{m,k,t} + \alpha}{\sum_{i=1}^K n_{m,i} + K\alpha} \quad (7)$$

$$\psi_{k,w} = \frac{n_{k,w} + G_{m,t=w} + \beta}{\sum_{i=1}^V n_{k,i} + V\beta} \quad (8)$$

其中, $\sum_{t=1}^T G_{m,k,t}$ 表示文档 m 中主题 k 下所有特征词权重值之和; $G_{m,t=w}$ 表示主题 k 下的词 w , 若是主题特征词, 其值为该词对应的权重值, 否则为 0。两个公式中的分母在进行求和时, 同样也需要加上其相应的权重值。

已知文档集中“文化产业”、“世博会”、“金融机构”、“知识产权”、“长江三角洲”等词语都属于文档集中文档的主题特征词语, 它们在主题建模过程中应该起着很重要的作用。为了增加这些词语的重要性, 在进行初步采样时, 首先把这些词作为主题特征词进行标记加入主题引导词库中, 然后计算它们各自在每篇文章的权重值, 最后在进行 Gibbs 采样的过程中, 扫描到这些词语后, 在统计频率时加入它们对应的权重值, 即式 (7) 和式 (8) 的实现过程。

按照改进后的实现步骤进行相应的实验得到建模结果, 其中包含了主题特征词的部分主题分布, 与表 1 中展示的结果相比, 那些主题引导词库中的主题特征词在主题词中的位序明显有所提高。部分主题分布对比的情况如下:

主题 3 “文化产业”、“世博会”分别上升了 1 个位序和 3 个位序;

主题 4 “金融机构”上升了 5 个位序;

主题 5 “知识产权”上升了 7 个位序;

主题 6 “长江三角洲”上升了 4 个位序。

通过上述的建模过程, 可以促使采样的主题分布偏向主题特征词的方向, 同时那些常见高频词的影响就会有所降低, 最终使文档的主题建模更加准确。

4 实验

实验主要通过应用 LDA 模型改进建模方法前后两种情况的对比进行。

4.1 实验分析

实验中文档集使用的是社科文献领域类的专题文献。对于文档集的数量大小, 依次是 100 篇、1 000 篇、5 000 篇、10 000 篇。 α 的默认值是 $1/K$ (K 是主题数量, 取值为 20); β 一般设置为 0.02; δ 默认取 0.02; 吉布斯采样的迭代次数一般设置为 2 000。

为了保证实验的对比性, 其他参数都是相同的。即相同文档集下, α 、 β 、主题数 K , 以及迭代次数都是一样的。

对相同的文档主题特征词增加权重值前后在主题建模后的模型结果中的情况进行对比。图 3 展示了主题特征词 (t1: 文化产业, t2: 世博会, t3: 金融机构, t4: 知识产权, t5: 长江三角洲) 在对应主题中的概率值的变化。结果表明, 这些主题特征词增加权重后, 其概率值明显有所增加, 即它们在建模过程中对主题分布的影响有所增强。

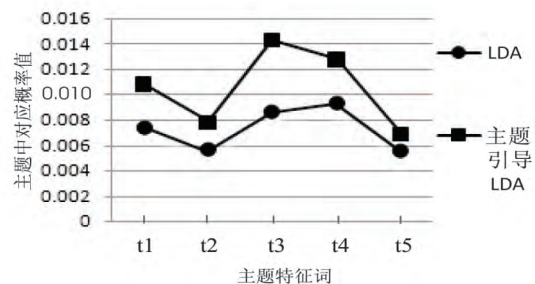


图 3 主题特征词概率值对比

两种不同建模方法得到的主题模型中, 对同一个主题下的主题词之间的相似度, 即主题的明确度进行对比, 相似度越高, 文档主题提取的越明确。图 4 展示了两种模型结果中部分主题的主题词相似度对比情况。

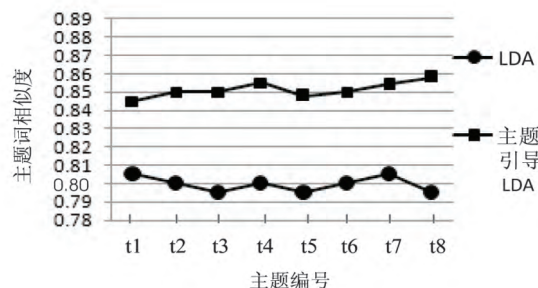


图 4 主题的主题词相似度对比

图5为在两种不同方法下建模后文档主题之间的相似度的对比情况。结果表明,主题之间相似度越低,文档集主题提取的类别越准确。

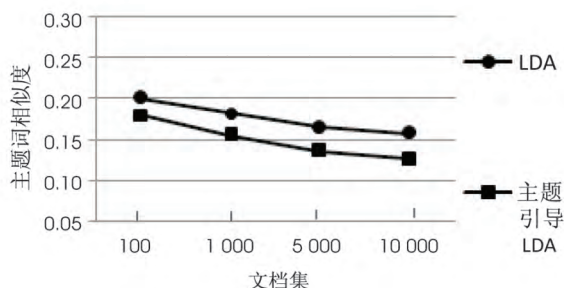


图5 主题之间的相似度对比

4.2 实验结果

通过对主题引导词的加权,提高主题引导词在文档中的重要性,从而影响文档和文档集的主题分布情况,最终使主题建模的结果更加符合文档集自身的主题分布特点。实验表明,在主题建模过程中增加文档集中主题特征词的权重进行主题建模的方法是行之有效的。

算法中对主题引导词加权时需要用到的引导参数 δ 依据经验选取了0.02。在具体的应用中,可以根据主题引导词对文档集的重要性进行适当改变, δ 值越大,引导词对主题的贡献率就越大,对主题的分布就越偏向该引导词。

5 结束语

提出一种针对社科文献领域类的文档集的主题建模方法。首先利用文档集中主题特征词处理得到主题引导词库,然后计算主题引导词权重并将其权重值增加到建模过程中,引导模型的主题分布,最后得到符合文档集自身主题特点的建模结果。

实验结果表明,该方法可以成功引导主题分布的情况,达到建模后的主题更加符合文档集本身主题分布特点的目的。

参考文献:

- [1] 王 昱.社科文献的特点、作用及省级社科文献资源建设[J].青海社会科学,1994(6):83-89.
- [2] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by latent semantic analysis[J]. Journal of the Association for Information Science and Technology, 1990, 41(6): 391-407.
- [3] DAN O. Probabilistic latent semantic analysis[C]//Proceed-

ings of uncertainty in artificial intelligence. [s.l.]: [s.n.], 1999: 289-296.

- [4] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.
- [5] 张志飞, 苗夺谦, 高 灿. 基于 LDA 主题模型的短文本分类方法[J]. 计算机应用, 2013, 33(6): 1587-1590.
- [6] 高 明, 金澈清, 钱卫宁, 等. 面向微博系统的实时个性化推荐[J]. 计算机学报, 2014, 37(4): 963-975.
- [7] YANO T, COHEN W W, SMITH N A. Predicting response to political blog posts with topic models[C]//Human language technologies: the 2009 conference of the north american chapter of the association for computational linguistics. [s.l.]: Association for Computational Linguistics, 2009: 477-485.
- [8] 张晓艳, 王 挺, 梁晓波. LDA 模型在话题追踪中的应用[J]. 计算机科学, 2011, 38(10A): 136-139.
- [9] 何锦群. LDA 在信息检索中的应用研究[D]. 天津: 天津理工大学, 2014.
- [10] 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.
- [11] 彭 云, 万常选, 江腾蛟, 等. 基于语义约束 LDA 的商品特征和情感词提取[J]. 软件学报, 2017, 28(3): 676-693.
- [12] 胡勇军, 江嘉欣, 常会友. 基于 LDA 高频词扩展的中文短文本分类[J]. 现代图书情报技术, 2013(6): 42-48.
- [13] 张小平, 周雪忠, 黄厚宽, 等. 一种改进的 LDA 主题模型[J]. 北京交通大学学报: 自然科学版, 2010, 34(2): 111-114.
- [14] 施乾坤. 基于 LDA 模型的文本主题挖掘和文本静态可视化的研究[D]. 南宁: 广西大学, 2013.
- [15] 倪丽萍, 刘小军, 马驰宇. 基于 LDA 模型和 AP 聚类的主题演化分析[J]. 计算机技术与发展, 2016, 26(12): 6-11.
- [16] WALLACH H. Topic modeling: beyond bag of words[C]//Proceedings of the 23rd international conference on machine learning. Pittsburgh, Pennsylvania: [s.n.], 2006.
- [17] WEI Xing, 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. New York: ACM, 2006: 178-185.
- [18] NEVADA L V. Fast collapsed Gibbs sampling for latent Dirichlet allocation [C]//Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. New York, USA: ACM, 2008: 569-577.
- [19] SALTON G. Introduction to modern information retrieval [M]. New York: McGraw-Hill Book Company, 1983.