DOI:10.16515/j.cnki.32-1722/n.2021.02.007

基于线性回归模型的单词加权LDA主题识别方法研究

邰 悦,葛 斌*

(安徽理工大学计算机科学与工程学院,安徽 淮南 232001)

摘 要:针对社会化标签系统下Web资源存在大量潜在知识以及资源之间存在着独立性的问题,提出一种基于线性回归模型的单词加权潜在狄利克雷分布(LDA)的主题识别方法。通过线性回归模型建立任意文本资源之间的拟合函数,使用拟合函数获取每个资源的权重值,解决资源之间存在独立同分布的问题,并对拟合函数的数据点进行加权操作,进而实现语料库中每个单词的加权,最终获得字典单词的权重系数。在单词加权基础上建立单词加权LDA模型,通过吉布斯采样对Web资源的潜在主题进行深入挖掘。实验结果表明,相比传统主题模型,新的单词加权LDA算法在Web资源上具有更好的主题识别效果。

关键词:线性回归模型;单词加权;LDA;吉布斯采样

中图分类号:TP399 **文献标识码**:A **文章编号**:1672-755X(2021)02-0039-07

Research on Word-weighted LDA Topic Recognition Method Based on Linear Regression Model

TAI Yue, GE Bin*

(Anhui University of Science and Technology, Huainan 232001, China)

Abstract: Aiming at the existence of a large amount of potential knowledge and the independence of resources in Web resources under the social tagging system, a word-weighted LDA(latent Dirichlet allocation) topic recognition method based on linear regression model is proposed. We establish a fitting function between arbitrary text resources through a linear regression model, use the fitting function to obtain the weight value of each resource to solve the independent and identical distribution characteristics in resources. And the weighting operation on the data points of the fitting function is used to achieve the weight of each word in the corpus, and finally obtain the weight coefficient of the dictionary word. A word-weighted LDA model is established on the basis of word weighting, and the potential topics of Web resources are deeply explored through Gibbs sampling. Experimental results show that the new recognition method has better topic recognition effects on Web resources than traditional topic models.

Key words: linear regression model; word weighting; LDA; Gibbs sampling

社会化标签系统是Web2.0的一个重要应用形式,为互联网用户提供一个网络资源的管理平台。社会化标签是互联网用户在社会化标签系统下自发产生的元数据,用户可以根据自身对网络资源的喜好,对

收稿日期:2021-04-02
基金项目:国家自然科学基金(51874003,61703005);安徽省自然科学基金(1808085MG221)
作者简介:邰悦(1995—),男,安徽合肥人,硕士研究生,主要从事数据挖掘与智能信息处理研究。
通信作者:葛斌(1973—),男,安徽安庆人,教授,博士,主要从事数据挖掘与信息安全研究。

网络资源加以评论并标注合适的标签。这些社会化标签中存在着大量潜在的语义信息，具有重要的研究与应用价值[1]。

近年来对社会化标签的研究热度依旧持续不断。例如 Shi 等[2]构造 UATM（user-based aggregation topic model）用于分析社交数据，分析用户的偏好和意图分布，利用 RNN（recurrent neural network）和 IDF（inverse document frequency）来构造权重，并提出折叠的吉布斯采样（Gibbs sampling）算法用于 UATM 模型推断。Li 等[3]提出一种基于共现的标签谱聚类方法，该方法直接利用标签共现关系来测定标签的相关性，也利用用户、资源和标签之间的三元关系，更好地解决标签资源的高维和稀疏问题。Indra 等[4]利用随机奇异值分解（randomized singular value decomposition，RSVD）对潜在语义索引进行预测，通过实验证明标签推荐效果在召回率、准确率等指标上有明显改进。由于社会化标签内容具有多元性和复杂性，Allam 等[5]提出了一种添加丰富内在动机的概念，将用户在社会化标签方面的行为作为主要的预测因素，将使用标签工具时的用户分为三种状态，并通过实验验证了该三维概念的理论模型。

对标签数据进行主题建模是研究社会化标签的重要环节，目前主题模型在文本数据分析中已被广泛应用。在 PLSA（probabilistic latent semantic analysis）模型[6]的基础上，Blei 等[7]对 PLSA 模型进行扩展并提出了 LDA（latent Dirichlet allocation）主题模型来对文档集合进行建模，从而发现文档中潜在的语义结构。LDA 是在 PLSA 的基础上引入了狄利克雷（Dirichlet）先验分布超参数，形成一个包含文档、主题和词的三层贝叶斯结构模型。蒋竞等[8]使用 LDA 模型对中文软件问答社区进行了主题研究分析。其他研究人员也在 LDA 的基础上对其进行改进和变形，如 Tian 等[9]采用舍入重参数技巧（rounded reparameterization trick，RRT），在 VAE-LDA（variational auto encoder-LDA）基础上，舍入重参数化 Dirichlet 分布并提出 RRT-VAE 模型。Li 等[10]在 L-LDA（labeled-LDA）[11]的基础上进一步扩展，提出 SL-LDA（supervised labeled-LDA）对社会化标签进行分类。Das[12]采用两种正则化主题模型促进了额外的共现信息，使学习的主题更加合理，并提出了 Gaussian LDA。也有研究人员使用深度学习技术和 LDA 相结合，对微博数据进行分析[13-14]。由于大多数主题模型没有考虑到深层的语义，研究人员通过提取概念和命名实体来丰富评论资源，提出 Concept-LDA 模型用于挖掘在线评论系统的情感分析[15]。

传统主题模型方法在进行数据处理时通常采用词袋模型，由于社会化标签资源是由用户自行产生，存在着自发性和独立性等特性，传统主题模型对社会化标签进行主题识别会降低主题识别精度和语义理解度。文本之间存在独立同分布特性，导致社会化标签资源之间没有明显的关联性，建立其潜在关联关系并进行资源和单词加权是提高主题识别效果的有效方法。

1 线性回归模型拟合资源关系

假设社会化标签资源集合中有 M 个资源以及对应标签，将每个社会化标签资源和标签视为一体，并视为文本文档，其资源集合表示为 $R=\{r_1,r_2,\cdots,r_i,\cdots,r_M\}$，其中 r_i 表示文本评论集合中的第 i 个评论资源，且每个社会化标签资源由文本评论单词资源及其对应的标签单词资源组成并用集合表示为 $r_i=\{c_1^i,c_2^i,\cdots,c_n^i,\cdots,c_V^i\}$，其中 c_n^i 为资源 i 在字典中第 n 个单词的编号，V 表示 M 个资源及其对应标签共涉及 V 个单词。由于资源之间没有明显关联关系，通过线性回归模型来建立文本之间的拟合曲线（图1），从而获得社会化标签资源之间的潜在关联，基本思想如下：

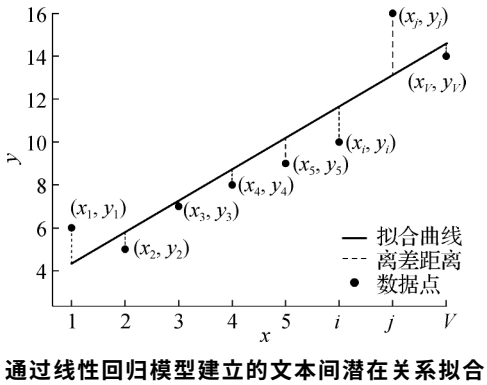

图1 通过线性回归模型建立的文本间潜在关系拟合曲线

以抽取不放回的方式抽取 M 个资源中的任意 2 个资源 r_i,r_j。假设 $r_i=\{c_1^i,c_2^i,\cdots,c_n^i,\cdots,c_V^i\}$，$r_j=\{c_1^j,c_2^j,\cdots,c_n^j,\cdots,c_V^j\}$，将 r_i 和 r_j 的单词合并为一个数据集。由于单词没有过多附属属性，只考虑词 c_n^i 在字典中是否存在。在实验过程中，文本 r_i 按字典序号递增排列，词 c_n^i 若存在即为在字典中的编号，若不存

在则为 0。由此形成 $d=\{(c_1^i,c_1^j),(c_2^i,c_2^j),\cdots,(c_V^i,c_V^j)\}$，由集合 d 形成数据集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_V,y_V)\}$。通过线性回归模型建立文本之间的拟合函数[16-17]，将数据集中自变量表示为矩阵 $\boldsymbol{X}$，因变量表示为矩阵 $\boldsymbol{Y}=\{y_1,y_2,\cdots,y_V\}$，建立 $f(\boldsymbol{x}_i)=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}_i+b$，使得 $f(\boldsymbol{x}_i)\approx y_i$，可由式(1)、式(2)获得线性回归模型式(3)。

$$\frac{\partial(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{A})^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{A})}{\partial\boldsymbol{A}}=2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{1}$$

$$\boldsymbol{A}^{*}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{2}$$

$$f(\boldsymbol{x}_i)=\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{3}$$

通过线性回归模型来获得 2 个文本的词向量的拟合函数，可以建立 2 个文本之间的拟合关系。

本文使用线性回归模型对文本资源进行拟合，将该拟合曲线(图 1)用于单词加权 LDA 算法的加权方法中，即通过拟合曲线获取每个数据点的离差向量，进而获得资源离差距离向量。

2　单词加权 LDA 算法模型描述

2.1　加权方法

由线性回归模型获得任意 2 个文本之间的单词拟合曲线(该曲线可以反映 2 个文本间的紧密程度)，并计算每一个数据点到拟合函数的离差距离，由此来形成任意 2 个文本之间的潜在关系。若拟合程度较高，说明 2 个文本间的关联性较强。通过加权方法建立资源的权重值，再将资源权重值作用到词频上得到单词的权值系数。计算出每个数据点到拟合函数数据点的离差距离 dp_i，得到向量 $\boldsymbol{DP}=[dp_1,dp_2,\cdots,dp_i,\cdots,dp_V]$。

$\boldsymbol{DR}=[dr_1,dr_2,\cdots,dr_i,\cdots,dr_V]$为资源离差距离向量，$dr_i$ 表示第 i 个文本数据点到拟合曲线平均离差距离。

$$dr_i=\frac{\|dp\|_1-\mathrm{sum}(dp^{\neg})}{V-\mathrm{sum}(dp^{\neg})} \tag{4}$$

$dp^{\neg}$表示离差距离为 0 的标量，$\mathrm{sum}(dp^{\neg})$为离差距离为 0 的标量个数总和，V 为字典词数。每个资源的权值系数为 γ，资源权值向量为

$$\boldsymbol{\gamma}=[\gamma_1,\gamma_2,\cdots,\gamma_i,\cdots,\gamma_M] \tag{5}$$

其中 $\gamma_i=\dfrac{dr_i}{\|\gamma\|_\infty}$表示资源 i 的权值，如果 i 和 j 是经过不放回采样得到的一组数据，则 $\gamma_i=\gamma_j$。M 个资源中涉及 V 个特征词，即字典数为 V。其频次向量为式(6)：

$$\boldsymbol{RF}_V(m)=[v_1,rf(v_1,m);v_2,rf(v_2,m);\cdots;v_i,rf(v_i,m);\cdots;v_V,rf(v_V,m)] \tag{6}$$

其中 $rf(v_i,m)$表示资源 m 中特征词 v_i 出现的频次。资源 m 的权重值为 $\gamma(m)$，资源特征词的加权表示形式为式(7)：

$$\boldsymbol{gRF}_V(m)=[v_1,grf(v_1,m);v_2,grf(v_2,m);\cdots;v_i,grf(v_i,m);\cdots;v_V,grf(v_V,m)] \tag{7}$$

其中

$$grf(v_i,m)=\mathrm{round}\left(\frac{\gamma(m)rf(v_i,m)}{\|\boldsymbol{\gamma}\|_{-\infty}}\right) \tag{8}$$

表示资源 m 中特征词 v_i 加权出现的频次。M 个资源中 V 个特征词的每个单词权重值向量，即加权词向量为

$$\boldsymbol{w}_v=[v_1,wgf(v_1);v_2,wgf(v_2);\cdots;v_i,wgf(v_i);\cdots;v_V,wgf(v_V)] \tag{9}$$

其中

$$wgf(v_i)=\sum_{m=1}^{M}grf(v_i,m) \tag{10}$$

上述加权方法首先通过线性回归模型获得拟合曲线，其次对资源进行拟合从而获得资源离差距离向量 $\boldsymbol{DR}$，通过向量 $\boldsymbol{DR}$ 生成每个资源的权值系数 γ，将资源权重值作用于资源单词的频次再对整个语料库

的单词进行累加，最终形成字典中每个单词的权重值，形成单词权值向量 $w=[w_1,w_2,\cdots,w_v]$。将 w 作用于 LDA 模型算法中，形成单词加权 LDA 模型。

2.2 单词加权 LDA 模型

由线性回归模型建立文本间的拟合关系，并通过加权方法获得词典的单词权值向量 w。已知社会化标签系统中有 M 个评论资源和对应的标签资源，假设存在 K 个潜在主题，字典词数为 V。单词加权 LDA 模型生成过程如下：

每个社会化标签资源都有各自的主题分布，该主题分布是多项分布，并且服从先验参数为 α 的 Dirichlet 分布，即对某一资源从 $Dir(\alpha)$ 中采样该资源的隐含主题。

每个潜在主题涉及各自的单词分布，且该单词分布为多项分布，该多项分布为先验参数 β 的 Dirichlet 分布，记为 $Dir(\beta)$。现基于线性回归模型生成拟合曲线，经过加权方法得到单词权值向量 w。w 是维度为 V 的向量，以高维先验参数作用于单词加权 LDA 模型，并且服从 $Dir(w)$ 分布。w 与参数 α 和 β 不同的是在采样过程中，传统 LDA 往往将 α 和 β 标量化并作用于 Dirichlet 分布。从 Dirichlet 先验分布 $Dir(w)$ 和 $Dir(\beta)$ 中采样潜在主题生成词分布参数 λ 和 φ。从参数为 θ 的多项式分布 $Mult(\theta)$ 中进行采样，采样出隐含主题 z，并从 z 所决定的多项式分布 $Mult(\varphi)$ 和 $Mult(\lambda)$ 中采样一个单词 W（图 2）。

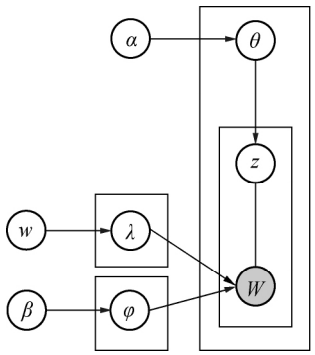

图 2 单词加权 LDA 模型

在单词加权 LDA 模型中，采用 Gibbs 采样方法来进行参数估计。Gibbs 采样的工作原理是根据潜在变量的后验分布生成样本，通过大量样本学习来训练模型和消除噪声。对于单词加权 LDA 中的每个单词对应的隐含主题，在每次迭代学习中，更新单词的隐含主题并保持其他主题分配不变。基于单词加权 LDA 的 Gibbs 采样主题和词分布如下：

$$\varphi_{i,j}=\frac{n_j^{(i)}+\beta_i}{\sum_{i=1}^{V}(n_j^{(i)}+\beta_i)},\quad \lambda_{i,j}=\frac{n_j^{(i)}+w_i}{\sum_{i=1}^{V}(n_j^{(i)}+w_i)},\quad \theta_{m,j}=\frac{n_m^{(j)}+\alpha_j}{\sum_{j=1}^{K}(n_m^{(j)}+\alpha_j)} \tag{11}$$

$$\begin{aligned}
&p(\boldsymbol{\theta}\mid \boldsymbol{z}_m,\boldsymbol{\alpha})=\frac{1}{Z_{\theta_m}}\prod_{n=1}^{N_m}p(z_{m,n}\mid\boldsymbol{\theta}_m)\cdot p(\boldsymbol{\theta}_m\mid\boldsymbol{\alpha})=Dir(\boldsymbol{\theta}\mid\boldsymbol{n}_m+\boldsymbol{\alpha})\\
&p(\boldsymbol{\lambda}\mid \boldsymbol{z},\boldsymbol{W},\boldsymbol{w})=\frac{1}{Z_{\lambda_k}}\prod_{\{i;z_i=k\}}p(W_i\mid\boldsymbol{\lambda}_k)\cdot p(\boldsymbol{\lambda}_k\mid\boldsymbol{w})=Dir(\boldsymbol{\lambda}\mid\boldsymbol{n}_k+\boldsymbol{w})\\
&p(\boldsymbol{\varphi}\mid \boldsymbol{z},\boldsymbol{W},\boldsymbol{\beta})=\frac{1}{Z_{\varphi_k}}\prod_{\{i;z_i=k\}}p(W_i\mid\boldsymbol{\varphi}_k)\cdot p(\boldsymbol{\varphi}_k\mid\boldsymbol{\beta})=Dir(\boldsymbol{\varphi}\mid\boldsymbol{n}_k+\boldsymbol{\beta})
\end{aligned} \tag{12}$$

其中 $n_j^{(i)}$ 表示词 i 分配主题 j 的次数，$n_m^{(j)}$ 表示主题 j 分配给文档 m 的次数。通过 Gibbs 采样进行参数估计，从加权后的语料库中学习 K 个主题并且采样出每个主题的词概率分布。

依据式(11)、式(12)采用单词加权 LDA 的 Gibbs 采样算法，从加权后的单词中学习得到资源的 K 个隐含主题以及所有特征词分别在 K 个隐含主题上的概率分布，可以估计单词加权 LDA 中“资源-主题”“主题-词”两语关系的概率。单词加权 LDA 在对单词加权的基础上挖掘资源的隐含主题，隐含主题从抽象的层面概括资源的词特征。对加权单词使用 Gibbs 采样，经过迭代学习，最终的采样分布结果逐渐收敛。基于单词加权 LDA 模型综合利用资源的内容信息，可以使学习到的主题能更好地表达资源的语义。

3 实验及性能分析

3.1 评价方法

本文实验采用公开数据集 CiteULike 和 Twitter，各采用 4 000 条数据，数据集的统计信息如表 1 所示。在模型进行训练前将所有数据集进行预处理，包括清洗掉停止词、无效标点符号、乱码符号等无效字符和无意义单词，通过预处理操作后形成最终的实验数据。

表 1　实验数据集的统计信息

数据集	资源数	词数	总词频数	单词频率	资源平均词频数
CiteULike	4 000	11 560	68 473	5.9	17.1
Twitter	4 000	12 038	33 216	2.8	8.3

本文采用常用的两种主题模型评价方法，包括主题间平均相似度指数[18]和困惑度指数[19]。主题间平均相似度越小，说明模型生成效果越好，主题表达能力越强。主题间平均相似度为：

$$Avgcorre = \frac{2}{K(K-1)}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} corre(\varphi_i,\varphi_j) \tag{13}$$

其中 $corre(\varphi_i,\varphi_j) = \dfrac{\sum_{v=1}^{V}\varphi_{vi}\varphi_{vj}}{\sqrt{\sum_{v=1}^{V}(\varphi_{vi})^2\sum_{v=1}^{V}(\varphi_{vj})^2}}$ 表示 2 个主题间的相似度。

困惑度评价方法是信息论中的一种测量方法，是评价主题模型的常用方法。困惑度指数越低则该模型的表达效果越好，说明该模型价值越高。困惑度评价方法为：

$$perplexity(M) = \exp\left\{-\frac{\sum_{m=1}^{M}\ln p(w_m)}{\sum_{m=1}^{M}N_m}\right\} \tag{14}$$

其中，M 为文本数，w_m 是数据集中的词数，N_m 是数据集中的总词数。

3.2　评价分析

本文分别在 CiteULike 和 Twitter 数据集上使用 PLSA、LDA、L-LDA、G-LDA 和单词加权 LDA 五种主题模型，并在两种评价方法上观察 5 个主题模型的主题识别效果。在实验过程中，在使用各个主题模型进行主题学习时，需要对先验参数 α 和 β 设定初始值。这些先验参数和主题数在实验过程中全部依据经验设定，设定 $\alpha=50/k$，$\beta=0.01$，主题数 $k=20,40,60,80,100,120,140,160,180,200$。

图 3 和图 4 展示了各主题模型在 CiteULike 上进行主题建模的情况，并使用困惑度指数和主题间平均相似度指数衡量主题识别效果。可以看出，在 CiteULike 数据集上，单词加权 LDA 具有较好的主题识别效果，当主题数 $k=200$ 时，单词加权 LDA 的主题识别效果达到最优。从图 3 可以看出，当 $k=180$ 时，LDA，L-LDA 达到最优主题识别效果，此时单词加权 LDA 的困惑度指数还未达到最优，但单词加权 LDA 模型的困惑度指数已经优于其他模型。图 4 显示，当 $k=200$ 时单词加权 LDA 的主题间平均相似度指标优于其他模型。单词加权 LDA 通过先验参数 w 作用于词采样，在词采样过程中单词更具有分布性，使得采样出来的主题词更具有合理性。从图 3 和图 4 可以得出 G-LDA 明显优于 L-LDA、LDA 和 PLSA，单词加权 LDA 和 G-LDA 在 CiteULike 数据集上的平均困惑度指数分别为 2 371 和 2 611，单词加权 LDA 相对于G-LDA 约降低 9%。单词加权 LDA 和 G-LDA 在 CiteULike 数据集上主题间平均相似度指数的平均值分别为 0.063 30 和 0.069 86，单词加权 LDA 相对于 G-LDA 约降低 9%。

图 5 和图 6 展示了 PLSA、LDA、L-LDA、G-LDA 和单词加权 LDA 在 Twitter 数据集上的主题识别情况，也采用两种评价方法对主题识别效果进行评价。从图 5 可以看出当主题数 $k=160$ 时，单词加权 LDA 主题识别效果达到最优，并且 LDA 模型在 $k=200$ 时的困惑度指数依旧高于单词加权 LDA 在主题数 $k=160$ 时的困惑度指数。从图 5 可以得到单词加权 LDA 和 G-LDA 在 Twitter 数据集上的平均困惑度指数分别为1 903 和 2 162，单词加权 LDA 相对于 G-LDA 约降低 12%。从图 6 可以得到单词加权 LDA 和 G-LDA 在 Twitter 数据集上主题间平均相似度指数的平均值分别为 0.277 8 和 0.313 7，单词加权 LDA 相对于 G-LDA 约降低 11%。总体上可以看出单词加权 LDA 具有较好的主题识别效果。图 6 中单词加权 LDA 模型 $k=140$ 时达到最优识别效果，尽管 $k=160$ 时出现波动，但总体上优于 G-LDA 与其他模型。

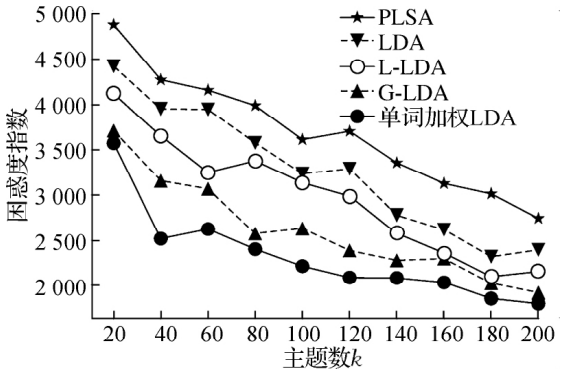

图 3 各方法在 CiteULike 数据集上的困惑度指数对比

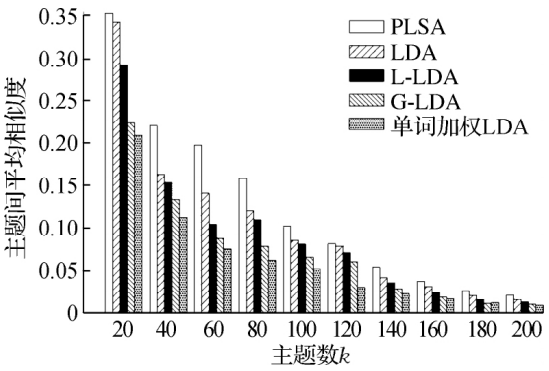

图 4 各方法在 CiteULike 数据集上的主题间平均相似度指数对比

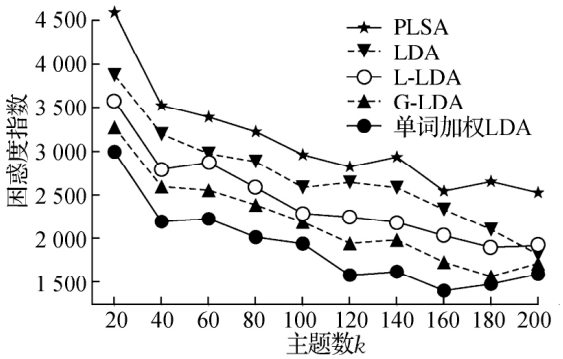

图 5 各方法在 Twitter 数据集上的困惑度指数对比

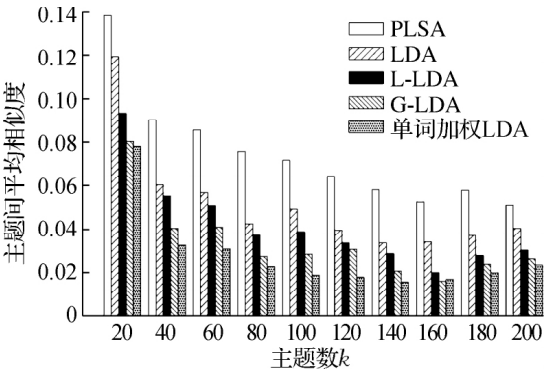

图 6 各方法在 Twitter 数据集上的主题间平均相似度指数对比

4 结 语

本文提出一种基于线性回归模型的单词加权 LDA 主题识别方法。首先通过线性回归模型对社会化标签资源进行拟合；然后使用加权方法对资源单词进行加权处理，通过单词加权建立单词加权 LDA 模型来对资源进行主题识别。这种主题识别方法可以对具有独立性的社会化标签文本建立潜在权重系数，并通过改进 LDA 对语料库进行深度主题挖掘。通过与其他主题模型在两种数据集上进行比较，单词加权 LDA 的各项评价指数均较好，但是对资源采样具有随机性以及高维与稀疏资源拟合不充分等问题的优化，还需要进一步研究。

参考文献：

[1] YAO J，WANG Y，ZHANG Y，et al. Joint latent Dirichlet allocation for social tags[J]. IEEE Transactions on Multimedia，2017，20(1)：224－237

[2] SHI L，SONG G，CHENG G，et al. A user-based aggregation topic model for understanding user's preference and intention in social network[J]. Neurocomputing，2020，413：1－13

[3] LI H，HU X，LIN Y，et al. A social tag clustering method based on common co-occurrence group similarity[J]. Frontiers of Information Technology & Electronic Engineering，2016，17(2)：122－134

[4] INDRA R，THANGARAJ M. An integrated recommender system using semantic Web with social tagging system[J]. International Journal on Semantic Web and Information Systems，2019，15(2)：47－67

[5] ALLAM H，BLIEMEL M，SPITERI L，et al. Applying a multi-dimensional hedonic concept of intrinsic motivation on social tagging tools：a theoretical model and empirical validation[J]. International Journal of Information Management，

2019，45：211－222

[6] HOFMANN T. Probabilistic latent semantic indexing[C]. California：Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval，1999：50－57

[7] BLEI D M，NG A Y，JORDAN M I. Latent Dirichlet allocation[J]. The Journal of Machine Learning Research，2003，3：993－1022

[8] 蒋竞，吕江枫，张莉. 中文软件问答社区主题分析研究[J]. 软件学报，2020，31(4)：1143－1161

[9] TIAN R，MAO Y，ZHANG R. Learning VAE-LDA models with rounded reparameterization trick[C]. Ponta Cana：Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing，2020：1315－1325

[10] LI X，OUYANG J，ZHOU X，et al. Supervised labeled latent Dirichlet allocation for document categorization[J]. Applied Intelligence，2015，42(3)：581－593

[11] RAMAGE D，HALL D，NALLAPATI R，et al. Labeled LDA：a supervised topic model for credit attribution in multi-labeled corpora[C]. Singapore：Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing，2009：248－256

[12] DAS R，ZAHEER M，DYER C. Gaussian LDA for topic models with word embeddings[C]. Beijing：Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processin，2015：795－804

[13] 黄佳佳，李鹏伟，彭敏，等. 基于深度学习的主题模型研究[J]. 计算机学报，2020，43(5)：827－855

[14] 陶玉婷，卓洋，张泽宇，等. 基于边界异类近邻关系构建的新特征提取方法[J]. 金陵科技学院学报，2018(3)：6－10

[15] EKINCI E，İlhan O S. Concept-LDA：incorporating Babelfy into LDA for aspect extraction[J]. Journal of Information Science，2020，46(3)：406－418

[16] 王惠文，叶明，GILBERT S. 多元线性回归模型的聚类分析方法研究[J]. 系统仿真学报，2009，21(22)：7048－7050

[17] 代亮，许宏科，陈婷，等. 基于 MapReduce 的多元线性回归预测模型[J]. 计算机应用，2014，34(7)：1862－1866

[18] 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，2010，34(2)：111－114

[19] LEE N，KIM E，KWON O. Combining TF-IDF and LDA to generate flexible communication for recommendation services by a humanoid robot[J]. Multimedia Tools and Applications，2018，77(4)：5043－5058

（责任编辑：湛　江）