基于LDA的多特征融合的短文本相似度计算

张小川 余林峰 张宜浩
(重庆理工大学计算机科学与工程学院 重庆 401320)

摘 要 近年来,LDA(Latent Dirichlet Allocation)主题模型通过挖掘文本的潜在语义主题进行文本表示,为短文本的相似度计算提供了新思路。针对短文本特征稀疏,应用LDA主题模型易导致文本相似度计算结果缺乏准确性的问题,提出了基于LDA的多特征融合的短文本相似度算法。该方法融合了主题相似度因子ST(Similarity Topic)和词语共现度因子CW(Co-occurrence Words),建立了联合相似度模型以规约不同ST区间下CW对ST产生的约束或补充条件,并最终权衡了准确性更高的相似度结果。对改进后的算法进行文本聚类实验,结果表明改进后的算法在F度量值上取得了一定程度的提升。

关键词 LDA,主题模型,短文本相似度,主题相似度,词语共现度
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.09.044

Multi-feature Fusion for Short Text Similarity Calculation Based on LDA

ZHANG Xiao-chuan YU Lin-feng ZHANG Yi-hao
(College of Computer Science and Engineering,Chongqing University of Technology,Chongqing 401320,China)

Abstract In recent years,latent dirichlet allocation(LDA)topic model provides a new idea for short text similarity calculation by mining the latent semantic themes of text. In view of the sparse features of short text,because the application of LDA theme model may easily lead to inaccurate results of similarity computation,this paper presented a calculation method based on LDA model combining similarity topics factor ST and co-occurrence words factor CW to establish union similarity model. In the protocol of different ST intervals,CW generates constraint or supplementary conditions to ST,and obtains higher accuracy of text similarity. A text clustering experiment was used to verify the method. The experimental results show that the proposed method gains a certain improvement of F measure value.

Keywords LDA,Topic model,Short text similarity,Similarity topics,Co-occurence words

1 引言

随着文本数据研究的不断深入,短文本作为互联网中广泛存在的一种文本数据,正逐步展现出强大的数据价值。如何对短文本进行有效挖掘和分析,成为一个研究热点。

短文本相似度计算是文本研究领域的基础工作。目前,相关计算算法试图挖掘短文本语义空间进行向量化表示,以解决传统模型对短文本泛化力差和特征表示困难的问题,将文本到词语的关系抽象到隐含语义空间,通过语义空间的相关性衡量文本相似度[1]。LDA是一种分析文本潜在语义空间的主题模型,该模型通过词项在文档级的共现信息来抽取语义相关的主题集合,并构建出文本的主题概率分布,通过文本主题概率计算主题相似度,由主题相似度映射文本相似度[2]。

主题相似度能够较好地概括文本语义相关度。相关度区别于相似度,相似度要求更高的准确性[3],涉及词语层和语义层两个层面的特征信息融合,完全依赖文本潜在语义表示的主题相似度,缺乏准确性。此外,短文本特征稀疏,LDA模型在计算过程中的主题聚焦性差,从而进一步影响了主题相似度的计算结果。词语层特征中,词语共现度计算是对文本相似特征的直接提取,认为文本间共现词语的概率越高,其相互关联越紧密,并且承载了一定的语义概念[4]。引入共现词语特征计算对短文本共现词集进行扩充和归并,得到的词语共现度将辅助主题相似度的计算。

针对现有LDA主题相似度方法的计算结果缺乏准确性的问题,本文提出了一种多特征融合的相似度计算方法。该方法从语义特征和词语特征两个层面进行文本表示,具体提出了选择扩充策略和联合相似度模型为基础的改进方法,实验表明该改进方法能显著提高文本相似度计算的准确性。

2 相关工作

2.1 LDA模型

LDA模型是一种引入全概率模型的文本主题表示方法[5],其核心是根据文本主题分布和主题词语分布的狄利克雷先验假设,结合词语样本信息计算文本后验主题分布的贝叶斯估算过程。模型可以对语料库 D 中的任意文本 m 建模,生成对应的主题概率分布 $\vec{\theta}_m=(z_{m,1},z_{m,2},\cdots,z_{m,n})$。模型推

到稿日期:2017-07-11 返修日期:2017-10-08 本文受国家自然科学基金(60443004),重庆市重大科技项目(cstc2013jcsf-jcssX0020),重庆市基础科学与前沿技术研究计划项目(cstc2015jcyjA40041)资助。

张小川(1965-),男,硕士,教授,主要研究方向为人工智能、计算机软件,E-mail:zxc@cqut.edu.cn;**余林峰**(1992-),男,硕士生,主要研究方向为人工智能,E-mail:517844894@qq.com(通信作者);**张宜浩**(1982-),男,博士,主要研究方向为自然语言处理。

导过程结合联合概率公式的描述如下：

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

式(1)描述了 LDA 模型生成文本的概率推导过程。其中，α 反映文本中隐含主题的先验分布，β 反映隐含主题下词的先验分布，$\vec{\theta}_m$ 表示文本 m 的主题概率分布，$\vec{w}_m$ 表示文本 m 的词语集合。LDA 的有向概率图表示如图 1 所示。

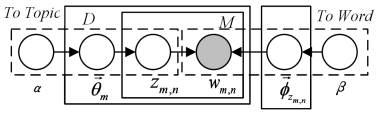

图 1　LDA 的有向概率图表示

Fig. 1　Representation of directed probability graph

图 1 中的 LDA 模型主要包括两个推导过程：

1)根据主题的先验分布参数 α 采样主题分布 $\vec{\theta}_m$，从主题分布 $\vec{\theta}_m$ 中取样生成一个主题 $z_{m,n}$ 的过程。概率公式如下：

$$p(z_{m,n}|\alpha)=p(z_{m,n}|\vec{\theta}_m)p(\vec{\theta}_m|\alpha) \quad (2)$$

2)根据主题词语的先验分布参数 β 采样每个主题对应的词语分布 $\vec{\phi}_{z_{m,n}}$，从词语分布 $\vec{\phi}_{z_{m,n}}$ 中采样最终生成词语 $w_{m,n}$ 的过程。概率公式如下：

$$p(w_{m,n}|z_{m,n},\beta)=p(w_{m,n}|\vec{\phi}_{z_{m,n}})p(\vec{\phi}_{z_{m,n}}|\beta) \quad (3)$$

联合概率公式模拟了随机生成一篇文本的一般性概率过程，并在过程中挖掘能够体现文本语义空间的主题分布。

2.2　Gibbs 抽样

构建 LDA 模型需要估算隐含参数 $\vec{\theta}_m$ 和 $\vec{\phi}_{z_{m,n}}$。本文采用 Gibbs 抽样算法[6]，估算过程可以看作文本生成的逆过程，即在给定文本集的情况下，通过参数估计得到隐含参数值。

Gibbs 抽样在于确定每个词语的主题，隐含参数便可以通过统计主题频数获得。因此，假定在排除当前词的主题 $\vec{z}_{-i}$ 分配的情况下，根据其他词的主题分配估计当前词分配各个主题的概率，计算公式如下：

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

其中，$z_i=k$ 表示词语 i 确定的主题 k；$-i$ 表示不包括词语 i 的其他主题集合；$n_{k,-i}^{(t)}$ 表示 k 主题中出现词语 t 的次数；$n_{m,-i}^{(t)}$ 表示文本 m 出现主题 k 的次数。假定每个词语的主题被确定，则可以按式(5)和式(6)估算 θ_{mk} 和 φ_{kt} 的值：

$$\theta_{mk}=\frac{n_{m,-i}^{(k)}+\alpha_k}{\sum_{k=1}^{K}(n_{m,-i}^{(k)}+\alpha_k)} \quad (5)$$

$$\varphi_{kt}=\frac{n_{k,-i}^{(t)}+\beta_t}{\sum_{t=1}^{V}(n_{k,-i}^{(t)}+\beta_t)} \quad (6)$$

其中，θ_{mk} 表示文本 m 中主题 k 的概率，φ_{kt} 表示主题 k 中词语 t 的概率，循环迭代计算文本 m 的主题分布 $\vec{\theta}_m$ 及语料的主题词语分布 $\vec{\phi}$。

2.3　基于 LDA 的相似度研究

LSA(Latent Semantic Analysis)通过奇异值分解方式概括文本潜在语义，未引入概率模型，语义空间的含义不明确，不利于文本表示[7]。pLSA(probabilitistic Latent Semantic Analysis)在此基础上用概率模型对每个变量的出现概率作出解释，但也面临概率模型不完备的问题[8]。LDA 引入先验分布的全概率模型，利用贝叶斯推断求解参数，更符合文本特性的需求。因此，国内外学者基于 LDA 模型进行文本表示，计算文本相似度，相关研究如下。

(1)基于原始特征的模型

张超等[9]针对原始词语特征中不同词性对文本相似度贡献的差异性，提出一种 PST_LDA 词性标注的狄利克雷模型，对原始词语特征建立了不同词性主题，通过不同词性词语的主题概率值计算相似度。张群等[10]基于 Word2Vec 训练词向量合成“词”粒度的文本向量，基于 LDA 主题向量合成“文本”粒度的文本向量，通过向量拼接构建词向量和 LDA 相融合的文本表示，进而计算相似度。Ramage 等[11]提出可扩展的局部监督学习模型(Label-LDA)，将类别信息融入 LDA 模型，保证了短文本在全部类别的隐含主题上进行协同分配，克服了传统 LDA 必须在单个类别中强制分配隐含主题的缺陷，有效提高了 Twitter 分类任务中相似度计算的准确度。

(2)基于特征扩充的模型

Phan 等[12]提出一种利用外部知识使得数据更相关的计算模型，从 WEB 中获取被称为一般性数据的大量文本信息，对短文本进行信息扩充，利用 LDA 描述扩充后的文本主题分布，从而计算主题相似度。吕超镇等[13]对数据降噪后，直接基于原始词项训练 LDA 模型，得到每个主题下的最优主题词分布概率，在文本高频主题下利用信息增益算法选取最优词扩充到文本中，再次训练扩充词语后的文本的主题向量，计算主题相似度。相对于 Phan 等提出的先扩充法，该方法可被称为后扩充法，加强了扩充词项的相关度，改善了主题相似度的计算效果。

上述方法过度依赖 LDA 模型，均着眼于建模后的主题特征，忽略了词语层特征。张群等将词向量作为文本补足特征，采用向量拼接的方法，未考虑向量叠加的影响，未见两者相互作用的具体关系，计算结果无法客观反映文本相似度。

3　改进的相似度计算方法

针对上述问题，本文算法从语料文本中提取 ST 和 CW 两个特征因子，ST 体现文本语义相关性，CW 体现文本共现词语量，同时建立两者相互作用的联合相似度模型，以权衡更为准确的相似度结果。本文算法的逻辑描述如图 2 所示。

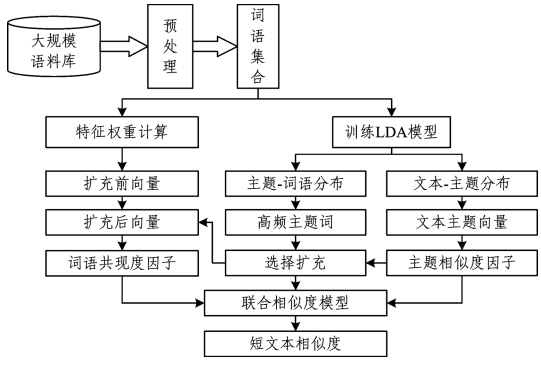

图 2　改进的相似度算法的框架

Fig. 2　Framework of improved similarity algorithm

多特征融合相似度算法由 4 部分组成:1)对语料库中的文本进行分词、去停用词预处理;2)训练 LDA 模型,输出文本-主题分布,计算 ST;3)基于选择扩充策略计算 CW;4)联合相似度模型,计算文本相似度。

3.1 主题相似度因子的提取

Gibbs 抽样可以估算语料中任意文本的主题概率分布向量,采用 JS 距离计算主题相似度。考虑到计算结果的对称性和有效范围为 0～1,JS 距离的计算式如下:

$$ST=D_{js}(\vec{\theta}_p,\vec{\theta}_q)=\frac{1}{2}(D_{kl}(\vec{\theta}_p,\frac{\vec{\theta}_p+\vec{\theta}_q}{2})+D_{kl}(\vec{\theta}_q,\frac{\vec{\theta}_p+\vec{\theta}_q}{2})) \tag{7}$$

其中,D_{kl} 表示 KL 距离,计算式如下:

$$D_{kl}(\vec{\theta}_p,\vec{\theta}_q)=\sum_{k=1}^{K}\theta_{p,k}\ln\frac{\theta_{p,k}}{\theta_{q,k}} \tag{8}$$

对于任意两个短文本 p 和 q,$\vec{\theta}_p$ 和 $\vec{\theta}_q$ 表示对应短文本主题分布向量,$\theta_{p,k}$ 表示短本 p 中主题 k 的概率值。

3.2 词语共现度因子的提取

短文本特征稀疏、共现词语少,易导致较低的词语共现度结果 Low-CW,在联合相似度的计算中 Low-CW 不利于相似度的提升。本文基于主题相似度阈值 C 提出选择扩充策略来计算 CW,对满足主题条件的文本进行特征扩充[14],扩充后的 CW 获得了一定比例的提升。扩充公式如下:

$$V_m'=\begin{cases}w_1,w_2,\cdots,w_i,s_1,s_2,\cdots,s_j, & ST<C\\ w_1,w_2,\cdots,w_i, & ST\geqslant C\end{cases} \tag{9}$$

其中,V_m' 表示语料中文本 m 基于 LDA 的扩展模型,w_1,w_2,…,w_i 为 TF-IDF 标记的文本原始特征,s_1,s_2,…,s_j 为选择扩充策略抽取的最优扩充特征词。抽取的扩充特征必须与文本有较强的相关性,文本主题分布中主题概率值大于阈值 F 的主题与文本的相关性强,该主题下概率值排在前 h 的词语对主题的表征性更强,此方式选取的高频词扩充到文本中将有助于提升 CW,其伪代码如算法 1 所示。

算法 1 选择扩充策略

输入:TF-IDF 标记文本向量集 $V=(\vec{V}_1,\vec{V}_2,\cdots,\vec{V}_M)$,LDA 标记文本主题向量集 $Z=\{\vec{\theta}_1,\vec{\theta}_2,\cdots,\vec{\theta}_M\}$,主题词语分布向量集合 $\phi=\{\vec{\phi}_1,\vec{\phi}_2,\cdots,\vec{\phi}_K\}$

输出:词语共现度结果 CW

过程:CW(V,Z,ϕ)

if 两待测文本主题相似度 $ST(\vec{\theta}_1,\vec{\theta}_2)\geqslant C$ then

  $CW=SIM(\vec{V}_1,\vec{V}_2)$计算文本向量余弦值;return

else

while 分别扩充两待测文本$(\vec{V}_m|m\in 1,2)$ do

  while 文本 m 的主题向量中大于 F 的主题 k do

    主题 k 词语分布$\vec{\phi}_k$ 排序后入栈 stack;

    for 抽取 stack 中 h 个排序在前的词语 do

      if !isempty(stack) then

        词语$\notin\vec{V}_m$,将其概率值加入高频词集 S;

      end if

    end for

  end while

  归一化 S,合并 S 和$\vec{V}_m$,生成$\vec{V}_m'$;

end while

$CW=SIM(\vec{V}_1',\vec{V}_2')$;

end if

3.3 联合相似度的计算

USIM(Union Similarity)联合相似度模型采用分段式联合线性加权的方法,ST 为基准值,$(1-ST)$为规约系数。规约不同区间下模型加成的补充或约束量,以达到权衡相似度准确值的效果。

(1)若 $ST<C$,则文本不相似,选择扩充策略扩充特征词。$USIM$ 表现为补充策略,$(1-ST)CW$ 为补充量,其中 CW 为规约量,ST 越低,CW 的补充力度越大,ST 越高,CW 的补充力度越小,公式如下:

$$ST+(1-ST)*CW,ST<C \tag{10}$$

(2)若 $ST\geqslant C$,则文本相似,未扩充特征词。$(CW-\lambda*ST)$为规约量,若 CW 达到 ST 的 λ 倍,则规约量正相关,否则负相关。

若 $CW/(\lambda*ST)<1$,$USIM$ 表现为约束策略,则规约量负相关,CW 越小,$(CW-\lambda*ST)$越大,约束力度越大,反之约束力度越小。

若 $CW/(\lambda*ST)>1$,$USIM$ 表现为补充策略,则规约量正相关,$\lambda*ST$ 越大,$(CW-\lambda*ST)$越小,补充力度越小,反之补充力度越大。

$$ST+(1-ST)(CW-\lambda*ST),ST\geqslant C \tag{11}$$

其中,λ 为拟合值。$\lambda\in(0,1)$服从 ST 和 CW 的数据分布,取值过程符合 $CW=\lambda*ST+b$ 线性拟合过程,如图 3 所示。

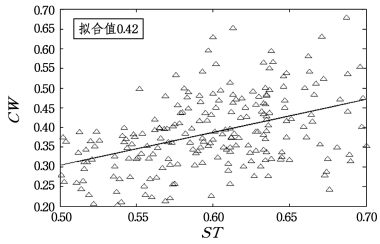

图 3 λ 线性拟合求解

Fig. 3 λ linear fitting solution

ST 和 CW 由本文实验数据计算得到,在此数据分布下 λ 的拟合值为 0.42。不难看出 $\lambda\propto CW$,CW 随 ST 线性增加,λ 的取值也应随之增大,因此 λ 的取值完全依赖于数据拟合得到。

综上所述,联合相似度模型的公式表述如下:

$$USIM=\begin{cases}ST+(1-ST)*CW, & ST<C\\ ST+(1-ST)(CW-\lambda*ST), & ST\geqslant C\end{cases} \tag{12}$$

为验证联合相似度模型的有效性,图示化联合相似度模型的实验效果。图 4 和图 5 分别给出了不同规约策略下 $UISM$ 可能达到的结果分布。

图 4 中 $ST<C$ 时表现为补充策略的三维图形,图形呈线性递增趋势,ST 在经过 CW 补充后取得了较好的补充效果,相似文本的比重较大。图 5 中 $ST\geqslant C$ 时表现为补充或约束策略的三维分段图形,其中 ST 不变,依据 $CW/(\lambda*ST)$的不同取值,若 CW 低于 $\lambda*ST$,则 CW 越低,曲面斜率越高,模型约束力度越大,$USIM$ 占据低取值结果分布;若 CW 高于 $\lambda*$

ST，则 CW 越大，曲面斜率越低，模型补充力度越小，$USIM$ 取值分布趋于平缓。

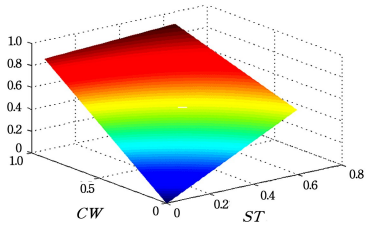

图 4　$ST<C$ 时联合相似度的变化

Fig. 4　Change of joint similarity when $ST<C$

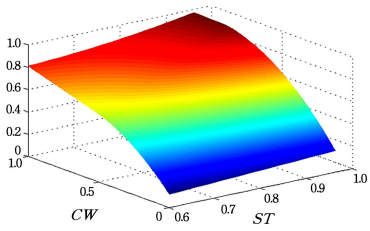

图 5　$ST\geqslant C$ 时联合相似度的变化

Fig. 5　Change of joint similarity when $ST\geqslant C$

4　实验分析

4.1　数据来源

实验所需短文本数据源于新浪微博网，通过 Python 爬取有明显类别特征的 1200 个用户微博，包括电影、美食、动漫、娱乐、星座和军事 6 大类话题，并依次对各类别数据编号。爬取的数据带有噪声字符，经过数据清洗操作过滤掉噪声字符，提取其中字数在 120～140 之间的文本数据作为实验语料，共计 10232 条，选取每个类别下 90% 的语料为训练语料，10% 的语料为测试语料。

4.2　评估指标

本文采用聚类算法中的 F 度量值评估各种相似度算法的效果，采用共现度方差 η 观察扩充前后词语共现度提高的百分比。其中，F 度量值是一种概括分类准确率 P 和召回率 R 的评测结果。F 度量值越大，相似度计算结果越准确；F 值越小，相似度计算结果的准确性越低。η 越大，相似度扩充效果越明显，反之同理。

$$\eta=\frac{\sum_{u=1}^{U}(CW_{ju}-C\overline{W})^2}{U} \tag{13}$$

$$P(P_j,G_i)=\frac{|P_j\cap G_i|}{|G_i|},R(P_j,G_i)=\frac{|P_j\cap G_i|}{|P_j|} \tag{14}$$

$$F(P_j,G_i)=\frac{2\cdot P(P_j,G_i)\cdot R(P_j,G_i)}{P(P_j,G_i)+R(P_j,G_i)} \tag{15}$$

其中，CW_{ju} 表示第 j 个人工标注类的第 u 对文本词语的相似度，U 为抽取文本对的数目，P_j 表示第 j 个人工标注的分类簇的文本数，G_i 表示第 i 个聚类簇的文本数，$P_j\cap G_i$ 表示同时存在分类簇 j 和聚类簇 i 的文本数。

4.3　实验结果及分析

4.3.1　参数估计及主题划分数目

LDA 模型的初始化参数配比（以下参数为多次实验的经验值配）如表 1 所列。

表 1　LDA 模型的初始化参数

Table 1　Initialization parameters of LDA model

α	β	Niters Gibbs
$50/K$	0.01	1000

表 1 中未知主题数 K 采用逐步递增的方式寻优，K 值区间为[10,100]，以 10 为间隔，递增寻优。聚类 F 值如图 6 所示。

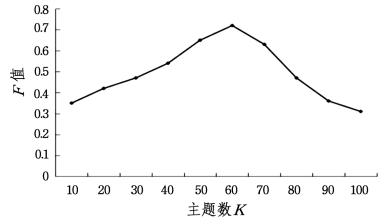

图 6　不同主题数下 F 值的变化

Fig. 6　Change of F value of different numbers of themes

显然，$K=60$ 时，F 值最高，聚类效果最佳，主题数 K 最优。

4.3.2　选择扩充策略的有效性验证

本实验验证选择扩充策略对词语共现度因子的提升效果，对比特征扩充前的词语共现度方差与扩充后的词语共现度方差（其中词语的平均相似度 $\overline{CW}$ 不变）的波动变化。方差波动效果如图 7 所示。

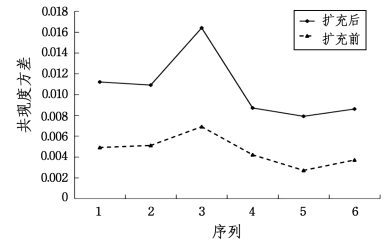

图 7　特征词扩充前后相似度方差的变化

Fig. 7　Variance of similarity caused by expansion of feature word

从图 7 可以明显看出，特征扩充后词语共现度方差均有上升，范围为 0.03%～0.26%，同时方差每提高 0.2%，CW 整体提高 10%，动漫类提高 0.26%，CW 提高 13%，扩充效果明显。

同时，测试选择扩充策略对文本相似度的影响。所有文本扩充策略（策略 1）、所有文本未扩充策略（策略 2）、选择扩充策略（策略 3）之间的聚类 F 值比较如表 2 所列。

表 2　不同扩充策略下聚类 F 值的比较

Table 2　Comparison of F values under different expansion strategies

（单位：%）

类别	F 值		
	策略 1	策略 2	策略 3
电影	0.54	0.47	0.63
美食	0.49	0.43	0.64
动漫	0.57	0.61	0.71
娱乐	0.52	0.43	0.63
星座	0.58	0.59	0.67
军事	0.52	0.61	0.69

下面对不同策略下的文本聚类 F 值进行分析。从整体上看，策略 3 的 F 值明显高于不采用策略 3（即采用策略 1 和策略 2）时的 F 值，聚类效果更理想，相似度计算更准确。这是因为策略 3 扩充了最优特征词向量，提高了 CW 的计算结

果；而其他策略对所有扩充情况一概而论，产生的词向量不具有适应性，计算 CW 时存在偏差，因此策略 3 更符合实际情况。

4.3.3 联合相似度模型的有效性验证

图 8 为不同 ST 取值区间下，LDA 主题相似度算法（方法 1）与本文改进算法（方法 2）的相似度值比较结果，其中 C 取值 50%（经验配比）。

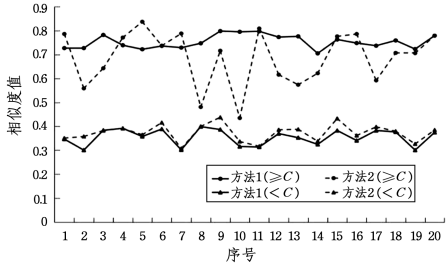

图 8 不同 ST 区间下相似度的改进对比

Fig. 8 Comparison of similarity improvement under different ST intervals

图 8 中，实线为方法 1 在不同 ST 区间下的基准值，当 $ST<C$ 时，方法 2 较基准值向上小幅波动，表明模型为补充效果；而当 $ST\geqslant C$ 时，方法 2 绕基准值上下波动，高于该值时模型表现为补充效果，低于该值时模型表现为约束效果。以上进一步表明，联合相似度模型确实影响了相似度值。

4.3.4 3 种相似度方法的聚类 F 值比较

为验证本文改进算法较其他相似度算法对性能提升的有效性，对 VSM 相似度算法、LDA 主题相似度算法和本文改进算法的聚类 F 进行了对比实验，结果如图 9 所示。

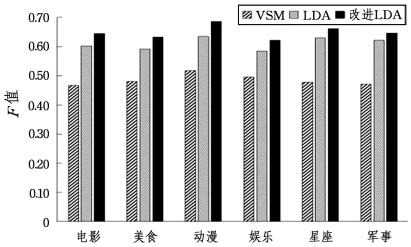

图 9 3 种相似度算法计算结果的比较

Fig. 9 Comparison of calculation results for three similarity algorithms

由图 9 可知，本文改进 LDA 算法比 LDA 主题相似度算法的 F 值提高了 3.8%，验证了改进算法的有效性；VSM 算法相比 LDA 主题相似度算法，F 值降低了 12.6%，相比本文的改进 LDA 算法，F 值降低了 16.4%。因此，本文提出的算法不论是在算法可行性还是实现效果上，均明显优于其他方法。

结束语 本文提出的多特征融合短文本相似度算法基于选择扩充策略提取词语共现度，并以此作为权衡主题相似度的辅助因子，建立了联合相似度模型，利用词语共现度合理规约主题相似度，得到了准确度更高的相似度结果。实验表明，本文算法较 VSM 和 LDA 主题相似度方法明显提高了文本相似度计算的准确度。下一步将考虑挖掘文本的其他特征，如词序、主题关联等，并将其整合后用于联合相似度计算中，以进一步改进基于 LDA 的相似度算法。

参 考 文 献

[1] CROFT D, COUPLAND S, SHELL J, et al. A Fast and Efficient Semantic Short Text Similarity Metric[C]// 2013 13th UK Workshop on Computational Intelligence. 2013: 221-227.

[2] CHEN P, YANG H, LV P, et al. Research on Text Similarity Based on LDA Model [J]. Computer Technology and Development, 2016, 26(4): 82-85. (in Chinese)
陈攀，杨浩，吕品，等. 基于 LDA 模型的文本相似度研究[J]. 计算机技术与发展，2016，26(4)：82-85.

[3] LIU H Z, XU D. Based Ontology Semantic Similarity and Correlation Computing Research [J]. Computer Science, 2012, 39(2): 8-13. (in Chinese)
刘宏哲，须德. 基于本体的语义相似度和相关度计算研究综述[J]. 计算机科学，2012，39(2)：8-13.

[4] CAO T, ZHOU L, ZHANG G X. A Text Similarity Calculation Based on Co-occurrence Words [J]. Computer Engineering and Science, 2007, 29(3): 52-53. (in Chinese)
曹恬，周丽，张国煊. 一种基于词共现的文本相似度计算[J]. 计算机工程与科学，2007，29(3)：52-53.

[5] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation [J]. the Journal of Machine Learning Research, 2003, 12(3): 993-1022.

[6] GibbsLDA++: A C/C++ Implementation of Latent Dirichlet Allocation(LDA) Using Gibbs Sampling for Parameter Estimation and Inference [EB/OL]. [2016-05-15]. https://sourceforge.net/projects/jgibblda/.

[7] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by Latent Semantic Analysis[J]. Journal of The American Society for Information Science, 1990, 41(6): 391-407.

[8] HOFMANN T. Probabilistic Latent Semantic Analysis[J]. Uncertainty in Artificial Intelligence, 1999, 56(8): 289-296.

[9] ZHANG C, CHEN L, LI Q. A PST_LDA Chinese Text Similarity Calculation Method [J]. Computer Application Research, 2016, 33(2): 375-377. (in Chinese)
张超，陈利，李琼. 一种 PST_LDA 中文文本相似度计算方法[J]. 计算机应用研究，2016，33(2)：375-377.

[10] ZHANG Q, WANG H J, WANG L W. Short Text Classification Method Based on Word Vector and LDA[J]. Modern Library and Information Technology, 2016, 32(12): 27-35. (in Chinese)
张群，王红军，王伦文. 词向量与 LDA 相融合的短文本分类方法[J]. 现代图书情报技术，2016，32(12)：27-35.

[11] RAMAGE D, DUMAIS S T, LIEBLING D J. Characterizing Microblogs with Topic Models[C]// International Conference on Weblogs and Social Media. Washington: ICWSM, 2010: 130-137.

[12] PHAN X H, NGUYEN L M, HORIGUCHI S. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections[C]// Proceedings of the 17th International Conference on World Wide Web. ACM, 2008: 91-100.

[13] LV C Z, JI D H, WU F F. Short Text Classification Based on LDA Feature Extension[J]. Computer Engineering and Applications, 2015, 51(4): 123-127. (in Chinese)
吕超镇，姬东鸿，吴飞飞. 基于 LDA 特征扩展的短文本分类[J]. 计算机工程与应用，2015，51(4)：123-127.

[14] HU Y J, JIANG J X, CHANG H Y. Chinese Short Text Classification Based on LDA High Frequency Word Expansion [J]. Modern Library and Information Technology, 2013, 16(6): 42-48. (in Chinese)
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，2013，16(6)：42-48.