混合词汇特征和LDA的语义相关度计算方法

肖 宝[1],李 璞[2,3],蒋运承[2]
XIAO Bao[1], LI Pu[2,3], JIANG Yuncheng[2]

1.钦州学院 电子与信息工程学院,广西 钦州 535011
2.华南师范大学 计算机学院,广州 510631
3.郑州轻工业学院 软件学院,郑州 450000
1.School of Electronics and Information Engineering, Qinzhou University, Qinzhou, Guangxi 535011, China
2.School of Computer Science, South China Normal University, Guangzhou 510631, China
3.Software Engineering College, Zhengzhou University of Light Industry, Zhengzhou 450000, China

XIAO Bao, LI Pu, JIANG Yuncheng. Combing lexical features and LDA for semantic relatedness measure. Computer Engineering and Applications, 2017, 53(12):152-157.

Abstract: Computing semantic relatedness in text documents is a key problem in many domains, for example, Natural Language Processing (NLP), Semantic Information Retrieval (SIR), etc. ESA (Explicit Semantic Analysis) for Wikipedia has received wide attention and applied mainly because of its simplicity and effectivity. However, use of ESA in semantic relatedness computation is inefficient due to its redundant concepts and high dimensionality. This paper presents a new technique based on LDA (Latent Dirichlet Allocation) and JSD (Jensen-Shannon Divergence) to computer semantic relatedness between text documents. The LDA is employed to reduce dimensionality and improve efficiency, and is used to build topic model probability vector from highly dimensional document matrix. Instead of cosine distance, JSD is used to compute semantic relatedness between documents. The results show that this technique based on LDA and JSD is more effective than ESA. Several benchmark test results have been presented to compare proposed technique with other methods. The results of experiment show that the proposed technique provides an increase of above 3% and 9% in Pearson correlation coefficient than ESA and LDA, respectively.

Key words: topic model; lexical features; Explicit Semantic Analysis(ESA); Latent Dirichlet Allocation(LDA); semantic relatedness measure

摘 要:文本语义相关度计算在自然语言处理、语义信息检索等方面起着重要作用,以Wikipedia为知识库,基于词汇特征的ESA(Explicit Semantic Analysis)因简单有效的特点在这些领域中受到学术界的广泛关注和应用。然而其语义相关度计算因为有大量冗余概念的参与变成了一种高维度、低效率的计算方式,同时也忽略了文本所属主题因素对语义相关度计算的作用。引入LDA(Latent Dirichlet Allocation)主题模型,对ESA返回的相关度较高的概念转换为模型的主题概率向量,从而达到降低维度和提高效率的目的;将JSD距离(Jensen-Shannon Divergence)替换余弦距离的测量方法,使得文本语义相关度计算更加合理和有效。最后对不同层次的数据集进行算法的测试评估,结果表明混合词汇特征和主题模型的语义相关度计算方法的皮尔逊相关系数比ESA和LDA分别高出3%和9%以上。

关键词:主题模型;词汇特征;显式语义分析(ESA);隐含狄利克雷分布(LDA);语义相关度计算
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.1606-0088

基金项目:国家自然科学基金(No.61272066);广州市科技计划项目(No.2014J4100031);广西高校中青年教师基础能力提升项目(No.KY2016LX431)。

作者简介:肖宝(1981—),男,讲师,主要研究方向:语义Web、机器学习;李璞(1983—),男,博士研究生,主要研究方向:Web语义分析、本体工程;蒋运承(1973—),男,博士,教授,博士生导师,主要研究方向:大数据语义分析、语义搜索、数据科学。

收稿日期:2016-06-06 **修回日期:**2016-07-25 **文章编号:**1002-8331(2017)12-0152-06
CNKI**网络优先出版:**2016-11-21, http://www.cnki.net/kcms/detail/11.2127.TP.20161121.2047.078.html

1　引言

语义相关度计算是自然语言处理的一个首要任务，也是信息检索中语义索引的一个重要技术支持[1-2]。对两个词“Apple”，“phone”进行是否相关以及相关程度如何的判断时，不同的人会有不同的答案，可能具有IT领域知识的知识分子与农民朋友会给出截然相反的结果。主要原因是背景知识的不同会对判断起到决定性的作用。又如单词“ontology”，它在计算机领域和哲学领域又是两个不同定义，人们可以通过该词所在的文章的上下文、所属主题(Topic)结合自身知识对该词进行认知。随着移动互联网技术的发展，计算机智能系统对词语、短文本作类似判断的需求越来越突出和重要。而计算机要像人一样具有思维并进行智能推理判断，首先其也必须具备相应丰富的领域知识[3]。目前计算机获取知识主要有两种方式，一种是利用搜索引擎的搜索结果[4-5]，另一种是采用外部的知识仓库如Wikipedia、WordNet、Open Directory Project[6-7]。后者则显得更加灵活和便于处理。在语义相关度计算中，类似WordNet知识库虽然具有结构化、语义定义完善的优点，但是构造和维护困难、知识覆盖范围小，使得用户数量和文章内容增长速度快、覆盖范围广的Wikipedia作为一个半结构化的知识库被更为广泛的应用[8-11]。Wikipedia的知识几乎无所不有，准确率几乎可以与百科全书相媲美[12]。在国内许多学者基于Wikipedia也做了相关度计算的研究[13-14]。然而这些研究基本上都是基于词汇特征以及Wikipedia类层次结构进行计算[15]，试图通过类间关系来寻找词汇的相关性。尽管都取得了不错的结果，但都需付出巨大的时间和空间代价来获得，文献[16]指出Wikipedia目前包含的类有644 950个，这些类数量巨大且结构组织层次没有优化，给相应的文章赋予类别时也过于随意，不能很好地反映文章(概念)与类别的信息，所以将之用于相关度计算的理由是不充分的。

有了领域知识，在进行语义相关度判断时还需要将词圈定在一定的主题(Topic)中。获取文本的主题，挖掘其背后隐含的语义模型被称之为主题建模(Topic modelling)，这已成为学术界研究的热点[6,17-19]。其中Blei[17]等人提出的LDA(Latent Dirichlet Allocation)已被文本分类、语义相关度计算等方面广泛应用[20-22]，目前LDA已成为了主题模型的标准。综上所述可以将机器模拟人类进行相关度计算过程表示如图1所示。

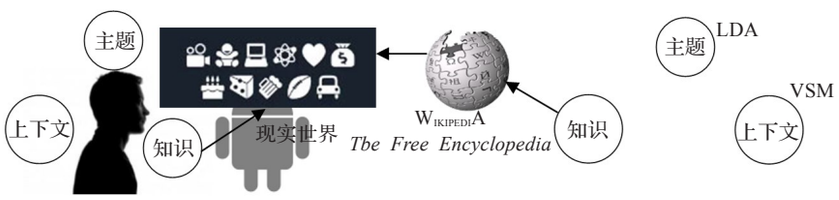

图1　机器模拟人类进行相关度计算过程

目前，许多工作都是基于Wikipedia的概念、类层次结构等信息或者单独对Wikipedia进行主题建模的研究应用[15]，为了结合词汇特征及其所属主题进行更加合理有效的语义相关度计算，本文在现有研究基础上提出一种混合词汇特征和主题模型的语义相关度算法，以提高相关度计算的精度和效率。

2　相关知识

2.1　ESA算法

ESA(Explicit Semantic Analysis)算法[3]是Evgeniy Gabrilovich基于Wikipedia设计的一个语义相关度的计算方法，该算法在自然语言处理领域的特殊贡献使得文献[3]获得了Journal of Artificial Intelligence Research的2014年度最佳论文奖。ESA首先为整个Wikipedia建立语义解释器，为所有的解释概念的文章中出现的词建立其TF-IDF权值倒排索引库，然后将要进行语义相关度比较的文本(词语)匹配到维基文章，并返回一个带有权重的维基百科概念向量，最后通过余弦测量方法计算相关度。图2是ESA算法的结构。

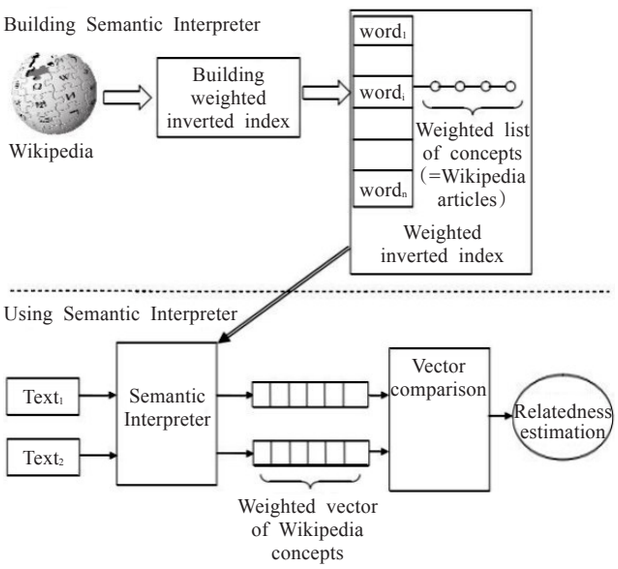

图2　ESA系统架构图

其算法步骤为：

(1)设Wikipedia的概念为 $C_1,C_2,\cdots,C_n$，解释相应概念文章为 $d_1,d_2,\cdots,d_n$。

(2)构造一个稀疏表 T，表的每一列对应于一个概念 C，每一行则对应于文档 d 中出现的某一个词。表中

的 $T[i,j]$ 的值是文档 d_j 中的词 t_i 的 TF-IDF 值。计算公式如下：

$$T[i,j]=tf(t_i,d_j)\cdot \ln\frac{n}{df_i} \tag{1}$$

其中词频计算公式如下：

$$tf(t_i,d_j)=\begin{cases}1+\ln count(t_i,d_j), \text{if } count(t_i,d_j)>0\\ 0, \text{otherwise}\end{cases} \tag{2}$$

df_i 指的是文档频率，表示包含词条 t_i 的文档数量，$df_i=\left|\{d_k:t_i\in d_k\}\right|$。为了消除文档长度对 $T[i,j]$ 的影响，对每一行的 $T[i,j]$ 的值利用公式(3)进行余弦归一化，公式(3)中的 r 是词条的数量。

$$T[i,j]\leftarrow\frac{T[i,j]}{\sum_{l=1}^{r}T[i,j]^2} \tag{3}$$

通过 $T[i,j]$ 就可以找到词与各个概念之间的关联强度。

(3)对输入的文本计算其 TF-IDF 值，作为权值，然后对每一个词查询倒排索引中计算好的概念空间向量，加权计算得到文本的表示向量。

设 $T=\{w_i|i=(1,2,\cdots,n)\}$，$\langle \boldsymbol{v}_i\rangle$ 是 w_i 的 TF-IDF 向量值，也是该词在输入文本中的权值。$\langle \boldsymbol{k}_i\rangle$ 是词 w_i 与 Wikipedia 的概念 c_j 的强度即是第二步计算好的权值，其中 $c_j\in c_1,c_2,\cdots,c_n$。计算 $\sum_{w_i\in T}\boldsymbol{v}_i\cdot\boldsymbol{k}_j$ 作为该词最终的向量值。

(4)对向量用余弦公式计算语义相关度。

2.2　LDA 算法

LDA[17]是一个基于概率的生成模型，它认为数据集包含有若干个主题，每个主题包含有若干个词，主题和词都是呈多项分布特征的。数据集中每一个文档可以看成是根据主题的概率选择了某个主题，再根据主题中包含的词概率选择某个词，重复步骤得到。LDA 的平滑版本模型如图3所示。

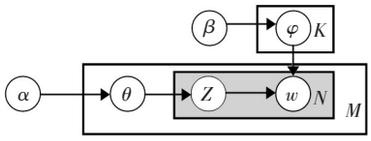

图3　LDA 模型图

在图中 M 是训练数据集中的文章总数，N 是文章的词的总个数，φ 是主题上的词分布，词 K 为主题总数，θ 是文章的主题分布，z 是每次生成文档词 w 时被选择的主题，因为一篇文档有多个主题，图中的灰色框表示选择词 w 及其相关的主题 z 的步骤重复进行 N 次，α 和 β 是两个超参数，分别表示每篇文档的主题分布的先验分布 Dirichlet 分布以及每个主题的词分布的先验分布 Dirichlet 分布。对于语料库中的每一篇文章可以用以下的生成过程得到：

(1)选择主题 k 上的词分布 $\varphi_k\sim Dirichlet(\beta)$，其中 $k=\{1,2,\cdots,K\}$；

(2)对文档集合 M 中的文档 m；

(2.1)选择文档 m 的主题 $\theta\sim Dirichlet(\alpha)$；

(2.2)对文档 m 中 N 个词 $w_n, n=\{1,2,\cdots,N\}$

① 选择一个主题 $z_n\sim multinomial(\theta)$；

② 选择一个词 $w_n|z_n\sim multinomial(\varphi_{z_n})$；

从图3以及生成过程可以得到文本中每个单词 w_i 的概率分布表示公式如公式(4)：

$$P(w_i)=\sum_{k=1}^{K}P(w_i|z_i=k)P(z_i=k), i=(1,2,\cdots,N) \tag{4}$$

文本的单词集合 w 与所属主题的联合概率分布计算公式如公式(5)：

$$P(w,z|\alpha,\beta)=\int P(z|\theta)P(\theta|\alpha)\mathrm{d}\theta\int P(w|z,\varphi)P(\varphi|\beta)\mathrm{d}\varphi \tag{5}$$

在模型中，文本主题概率分布和每个主题的词项概率分布是两组特别重要的参数，词项是语料库中的可以观测到的数据，而主题则是隐含在文本中看不到的变量，即 z,θ,φ 都是未知的隐含变量，为了得到主题相关的参数，Gibbs Sampling[23]是一种有效的方法，尤其是在语料数据很大时。该方法基于 MCMC(Markov Chain Monte Carlo)，这是一种通过不断迭代采样逼近主题后验概率函数的方法，迭代方式是每次选择概率向量的一个维度，给定其维度的变量值取样当前维度的值，不断迭代直到收敛输出待估计的参数为止，最后得到两个概率矩阵 $\boldsymbol{\theta}$(文档→主题) 和 $\boldsymbol{\varphi}$(主题→词)，前者表示每个文档中每个主题出现的概率，后者表示每个主题中每个单词的出现概率。

$$\boldsymbol{\theta}_{mat}=\begin{bmatrix}\theta_{1,1} & \cdots & \theta_{1,K}\\ \vdots & & \vdots\\ \theta_{M,1} & \cdots & \theta_{M,K}\end{bmatrix}$$

$$\boldsymbol{\varphi}_{mat}=\begin{bmatrix}\varphi_{1,1} & \cdots & \varphi_{1,W}\\ \vdots & & \vdots\\ \varphi_{K,1} & \cdots & \varphi_{K,W}\end{bmatrix}$$

其中矩阵中的每一项计算方法分别如公式(6)、公式(7)所示：

$$\theta_{m,k}=\frac{n_{n,k}+\alpha}{\sum_{i=1}^{K}n_{m,i}+K\alpha} \tag{6}$$

$$\varphi_{k,w}=\frac{n_{k,w}+\beta}{\sum_{i=1}^{V}n_{k,i}+V\beta} \tag{7}$$

3　混合词汇特征和LDA的语义相关度算法研究

从工程应用的角度上看，要计算词语或文本的语义相关度是否准确主要是其所依赖的上下文(context)、主题以及知识库。上下文的表达(语境)可以看作文本或词语周边的上下文中的其他词语，VSM[24](Vector Space Model)是一种表达词语的上下文信息最为经典的模

型。经过多年研究，VSM中的项有着多种复杂的上下文表达方式，如TF-IDF，互信息等，其中基于TF-IDF的VSM在文本挖掘、检索等方面应用最为广泛。ESA算法的基础正是基于此模型。然而ESA算法没有将主题考虑到算法中，其在语义相关度计算时首先获取待比较的文本与所有的Wikipedia的概念的相关度向量，然后进行向量余弦比较。算法存在两个问题：一是与文本相关度很低(甚至无关)的概念(文章)参与计算，浪费了时间和空间；二是因为Wikipedia的语料非常大，即使经过归一化的处理得到相关度也常常失真。为了提高相关度计算的效率本文提出一个新的算法，把主题模型LDA引入改造ESA算法。算法考虑了词汇特征和主题特性，其流程如图4。

整个流程可以分成三个部分：显式语义(explicit semantic)分析、主题模型LDA构建以及语义相关度计算。

3.1　显式语义分析

对要进行语义计算的文本进行分词、词干化，利用ESA的语义解释器(Semantic interpreter)通过词汇特征利用TF-IDF值找到与之相关度最大的概念$C_1,C_2,\cdots,C_n$，计算方法是2.1节的公式(1)~(3)。因为Wikipedia概念数量巨大，TF-IDF数值都比较小(不同的文本返回的值有可能相差一个数量级以上)，设置的合适阈值去限定返回概念数数比较困难，根据观察，前10个是最能反映显式语义的结果，表1是利用ESA对“computer science”语义解释后返回的前10条结果。

3.2　主题模型LDA构建

LDA模型构建的语料库是Wikipedia，Wikipedia中包含着许多保证日常运转、维护需要辅助文件，这些文件对模型的构建会起到负作用，所以在解释Wikipedia的dump文件时先去除命令空间为wikipedia、Category、File、Portal、Template、User、Help、Draft、MediaWiki的文件，然后去除语料库中的停用词并进行词干化。通过TF-IDF公式可以发现过少或过多出现的词对文本的分析作用比较微弱，为了降低LDA的学习成本，设定一个窗口去掉一部分词，本文把出现在不同文章的次数小于20及大于文章总数的10%词项全部去除。然后利用LDA结合Gibbs sampling通过2.2节的生成过程迭代得到概率矩阵$\boldsymbol{\theta}$(文档→主题)和$\boldsymbol{\varphi}$(主题→词)模型，设模型中的主题为$T=\{T_1,T_2,\cdots,T_n\}$，n是主题数。主题T_k包含的词项集合为$W=\{w_1,w_2,\cdots,w_m\}$，词项w_i属于主题T_j的概率为$P(w_i|T_j)$，则有：

$$\sum_{i=1}^{m}P(w_i|T_j)=1,\text{其中}j=\{1,2,\cdots,n\}$$

为了分析主题数量对LDA的影响，分别设定多个主题总数，表2是主题数量为3 000的第100、103、108的主题前8个词项。其中数字表示该主题下单词出现的概率。如“0.375*scenes”表示单词主题Topic100下“scenes”出现的概率为0.375。

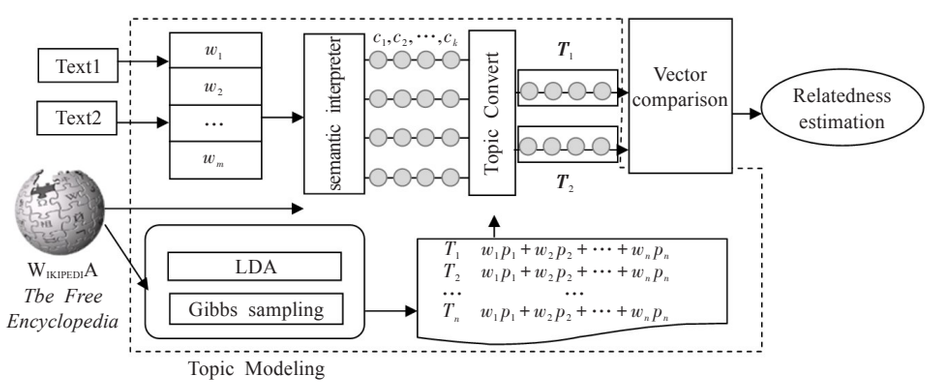

图4　混合词汇特征和主题模型的语义相关度计算流程

表1　语义解释器解释“computer science”结果

编号	Wikipedia ID	标题	得分
1	328784	Computer scientist	0.282 929 960 7
2	2443907	Department of Computer Science, University of Oxford	0.237 444 691 9
3	10433833	Computational mathematics	0.206 225 341 1
4	2701254	Bachelor of Computer Science	0.204 131 296 0
5	15832717	Computational statistics	0.182 799 390 4
6	659094	Quantum information science	0.174 376 585 1
7	361909	Computer lab	0.173 367 773 9
8	149353	Computational biology	0.169 682 063 6
9	12266843	Florida Atlantic University College of Engineering and Computer Science	0.169 389 417 0
10	4064368	David R. Cheriton School of Computer Science	0.164 578 201 4

表2 主题与词概率关系示例

Topic 100	Topic 103	Topic108
0.375*scenes	0.229*subjected	0.166*committed
0.280*camera	0.221*filling	0.164*injured
0.146*trademark	0.155*sheila	0.142*crimes
0.121*couples	0.111*puppets	0.105*convicted
0.038*harrington	0.084*condensed	0.074*jail
0.017*faire	0.077*acidic	0.064*murders
0.013*detractors	0.068*pigments	0.061*imprisonment
0.011*esque	0.017*alister	0.048*sentences

3.3 语义相关度计算

在上面的两个步骤中,LDA模型已将 $P(w,z|\alpha,\beta)$ 存储到文件中。设对文本Text经过语义解释返回Wikipedia的文章集合为 $D=\{d_1,d_2,\cdots,d_n\}$,词项为 $W=\{w_1,w_2,\cdots,w_m\}$,其中 $w_i\in D$ 。将训练好的LDA模型将 D 映射到二维主题空间中,得到主题向量 $\boldsymbol{T}=\{T_1,T_2,\cdots,T_n\}$ 。主题向量空间保存的是概率数据。测量概率分布之间的距离常用的方法有KL距离(Kullback-Leibler divergence)、卡方检验(Chi-Square),其中KL距离已成为概率向量相似度度量的标准。其公式如下:

$$D_{KL}(P,Q)=\sum_{j=1}^{T}P_j\ln\frac{P_j}{Q_j} \tag{8}$$

但是KL距离具有不对称性,即 $D_{KL}(P,Q)\neq D_{KL}(Q,P)$,所以本文采用KL距离的另一个变种JSD距离(Jensen-Shannon Divergence)替换ESA的余弦测量方法,计算公式如下:

$$JSD(P,Q)=\frac{1}{2}D(P,M)+\frac{1}{2}D(Q,M) \tag{9}$$

其中 $M=\frac{(P+Q)}{2}$ 。

4 实验设计与结果分析

评估语义相关度算法的有效性,通常是通过计算皮尔逊相关系数(Pearson correlation coefficient)得到的语义相关度值 X 与人工判断值 Y 相关程度。计算公式如下:

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{10}$$

4.1 语料与测试数据

目前Wikipedia的数据还在快速的增长,但是因为各种原因Wikipedia中文数据增长相对缓慢且覆盖范围小,所以本文使用最新英文Wikipedia作为语料库。测试集则采用WordSimilarity-353以及CLSS(Cross-Level Semantic Similarity)提供的数据集[25]的部分子集,前者包含353对词对,后者设计了四种不同语言层次比较类别集合:段落到句子(paragraph2sentence),句子到短语(sentence2phrase),短语到词(phrase2word),词到词义(word2sense),设定的相似度级别从0~4共5种,与WordSimilarity-353 一样都进行了人工评分。为了验证本文设计的算法的有效性,这里只选择CLSS数据集两个子集paragraph2sentence,sentence2phrase(各自包含500对测试条目)做测试和比较。

4.2 实验结果分析

对Wikipedia语料应用LDA建立的模型对语义相关度计算起着关键的作用,实验中设置Gibbs抽样迭代次数为1 000次,α 的值被研究[26]认为等于50/K(K 是主题数)时在不同的数据集上都能起到很好的作用,所以实验中根据主题数进行动态的设置,$\beta=0.001$ 。随着主题数量和LDA词典(特征词)的增加,模型所消耗的时间和空间复杂度急剧上升。实验机器硬件配置为:Intel® Core™ i3-4130 3.40 GHz CPU +16 GB内存。ESA算法采用文献[27]所提供的源代码进行测试,LDA和本文设计的语义相关度计算方法则使用了基于Python语言的机器学习库Gensim,主题数量 K 的值则分别设置为500,1 000,1 500,2 000,3 000共五种。表3展示了混合词汇特征和LDA的语义相关度计算方法在设置不同主题数情况下在各数据集中的皮尔逊相关系数计算结果。可以看到当主题 $K=1\,000$ 时,语义相关度计算的效果最好。

表4则是在主题 $K=1\,000$ 时,LDA和ESA以及本文提出的混合词汇特征和主题模型的语义相关度算法的皮尔逊相关系数结果。通过表4可以看到本文提出的相关度计算算法分别比ESA和LDA的皮尔逊相关系数高出3%~5%,9%~26%。可以发现,比较的文本包含的内容越多(意味着词项特征越多),计算的效果越好。其中的LDA在数据集WordSimilarity-353中值为0.000 000是因为实验中,每对词鲜有与模型中词典匹配,得到的值大部分都是0,在本实验中讨论该值没有太多的意义。

表3 混合词汇特征和LDA的语义相关度算法在不同数据集和不同主题数中的皮尔逊相关系数结果

数据集	主题数				
	500	1 000	1 500	2 000	3 000
WordSimilarity-353	0.369 109 8	0.397 820 5	0.346 075 2	0.345 929 9	0.343 073 1
sentence2phrase	0.745 437 7	0.764 189 0	0.763 760 2	0.753 374 1	0.757 713 7
paragraph2sentence	0.761 860 2	0.762 558 9	0.753 886 1	0.756 101 1	0.753 010 3

表4　三种语义相关度计算方法的皮尔逊相关系数比较结果

数据集	方法		
	ESA	LDA	混合词汇特征和LDA的语义相关度算法
WordSimilarity-353	0.345 724 9	0.000 000	0.397 820 5
sentence2phrase	0.734 680 6	0.500 359	0.764 189 0
paragraph2sentence	0.734 680 6	0.672 387	0.762 558 9

5　结束语

本文综合考察了LDA和ESA两种算法的原理和特点，提出了一种新的混合词汇特征和主题模型的语义相关度计算算法。该方法在ESA的基础上加入了LDA模型，使得计算机能像人类一样具有背景知识、能识别词项的所属的主题，在主题的限定范围下，结合VSM表达的上下文知识进行文本语义相关度的计算，结果表明这是一种更为合理有效的算法。该方法的提出可以为语义检索、自然语言信息处理等领域提供理论研究和应用基础。从实验结果上看LDA的主题对语义相关度计算有着较大的影响，在词汇特征较少的情况，效果较差，如何从巨大的语料库中学习到更加合理的主题数以及对词汇特征少的文本进行更为有效的语义相关度计算是本文的下一步工作。

参考文献：

[1] Zargayouna H.Contexte et sémantique pour une indexation de documents semi-structurés[C]//à paraîtredans ACM COnférence en Recherche Information et Applications，CORIA，2004：161-178.

[2] Formica A.Concept similarity in formal concept analysis：an information content approach[J].Knowlledge-Based Systems，2008，21(1)：80-87.

[3] Gabrilovich E，Markovitch S.Wikipedia-based semantic interpretation for natural language processing[J].Journal of Artificial Intelligence Research，2009，34：443-498.

[4] Bollegala D，Matsuo Y，Ishizuka M，et al.Measuring semantic similarity between words using web search engines[J].World Wide Web，2007，7：757-766.

[5] Yih W，Meek C.Improving similarity measures for short segments of text[C]//National Conference on Artificial Intelligence，2007，7(7)：1489-1494.

[6] Jin O，Liu N N，Zhao K，et al.Transferring topical knowledge from auxiliary long texts for short text clustering[C]//Proceedings of the 20th ACM International Conference on Information and Knowledge Management.ACM，2011：775-784.

[7] Chen M，Jin X，Shen D，et al.Short text classification improved by learning multi-granularity topics[C]//International Joint Conference on Artificial Intelligence，2011：1776-1781.

[8] Milne D.Computing semantic relatedness using wikipedia link structure[C]//Proceedings of the New Zealand Computer Science Research Student Conference，2007：1-8.

[9] Taieb M A H，Aouicha M B，Tmar M，et al.Wikipedia category graph and new intrinsic information content metric for word semantic relatedness measuring[M]//Data and Knowledge Engineering.Berlin Heidelberg：Springer，2012：128-140.

[10] Yeh E，Ramage D，Manning C D，et al.WikiWalk：random walks on Wikipedia for semantic relatedness[C]//Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing.Association for Computational Linguistics，2009：41-49.

[11] Jiang Y，Zhang X，Tang Y，et al.Feature-based approaches to semantic similarity assessment of concepts using Wikipedia[J].Information Processing and Management，2015，51(3)：215-234.

[12] Giles J.Internet encyclopaedias go head to head[J].Nature，2005，438(7070)：900-901.

[13] 孙琛琛，申德荣，单菁，等.WSR：一种基于维基百科结构信息的语义关联度计算算法[J].计算机学报，2012，35(11)：2361-2370.

[14] 王荣波，谌志群，周建政，等.基于Wikipedia的短文本语义相关度计算方法[J].计算机应用与软件，2015，32(1)：82-85.

[15] Taieb M A，Aouicha M B，Hamadou A B，et al.Computing semantic relatedness using Wikipedia features[J].Knowledge Based Systems，2013，50：260-278.

[16] Saif A，Aziz M J，Omar N，et al.Reducing explicit semantic representation vectors using latent dirichlet allocation[J].Knowledge Based Systems，2016，100：145-159.

[17] Blei D M，Ng A Y，Jordan M I，et al.Latent dirichlet allocation[J].Journal of Machine Learning Research，2003，3：993-1022.

[18] Lu Y，Mei Q，Zhai C，et al.Investigating task performance of probabilistic topic models：an empirical study of PLSA and LDA[J].Information Retrieval，2011，14(2)：178-203.

[19] Diao Q，Jiang J，Zhu F，et al.Finding bursty topics from microblogs[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics：Long Papers-Volume 1.Association for Computational Linguistics，2012：536-544.

（下转165页）