基于 LDA 的弱监督文本分类方法

张金瑞，柴玉梅，咎红英，高明磊

（郑州大学 信息工程学院，河南 郑州 450001）

摘 要：针对传统的文本分类方法需要大量人工标注好的训练数据，且数据标注的好坏会影响结果等问题，通过对 LDA 及其相关模型的研究，提出一种基于 LDA 的弱监督文本分类算法。无需人工标注训练数据，在处理文本时，引入词向量，保持文本中的词序，加入二元语法。实验结果表明，该方法节省了人力、物力，取得了较优效果。

关键词：文本分类；潜在狄利克雷分布；主题；词序；二元语法

中图法分类号：TP391.1 **文献标识号**：A **文章编号**：1000-7024（2017）01-0086-06

doi：10.16208/j.issn1000-7024.2017.01.017

Weakly supervised text classification method based on LDA

ZHANG Jin-rui, CHAI Yu-mei, ZAN Hong-ying, GAO Ming-lei

(School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract: To resolve problems that the traditional text classification methods need a lot of manually labeled training data and that the quality of the data influences the results, through the study of LDA and its related models, a weakly supervised text classification algorithm on the basis of LDA was presented. Manually labeled training data were no longer needed. Besides, when dealing with the text, the word vector was introduced, and the word order was maintained and the bigram grammar was joined. Experimental results show that when this approach reduces manpower and material resources, it also obtains better effects.

Key words: text classification; LDA; topic; word order; bigram grammar

0 引 言

目前，在文本分类领域，已有不少学者取得了一定的成果。Blei 和 Ng[1] 用 LDA 中的主题作为文本分类时的特征，但是在学习分类器时却用了带标注的文档集；Sangno 等[2] 将 LSA（latent semantic analysis）、PLSA（probabilistic latent semantic analysis）、LDA（latent dirichlet allocation）、CTM（correlated topic model）4 种方法在多组实验上进行了比较，并详细陈述了它们各自的特点和与之相适应的情况；Jun 和 Amr[3] 对 LDA 进行了改进，加入了最大熵算法，利用最大边界原理来训练有监督的主题模型，并估计和预测文档的主题表示，取得了不错的效果；Sanjeev 和 Rong[4] 通过对 LDA 的研究，引入了工具 NMF（nonnegative matrix factorization）来学习文档中的主题结构，在此领域也做出了一定的贡献；Animashree 等[5] 在 LDA 的基础上，利用统计中的三元或者四元模型，通过两个奇异值分解来训练文档中的主题模型，进而实现对文本的分类，以上几种方法大都需要标注的训练数据；Hingmire 等[6] 提出了 ClassifyLDA 方法，首先用 LDA 对无标注的训练数据生成一个主题模型，然后为每一个主题人工指定一个类别，再把文档中每个词的主题替换为它所指定的类别，这样就形成一个新的主题模型。给文档分类时，根据文档中最具代表性的主题是哪个类，此文档就归为哪个类。

在文本分类算法中，有监督的学习算法需要大量的标注好的训练数据，针对此问题，本文提出了一种基于 LDA 的文本分类算法 VB-LDA（latent Dirichlet allocation with vector and bigram），它不需要标注数据，只是在为主题选择类别时，不再人工指定，而是利用距离度量来计算当前

收稿日期：2015-10-27；**修订日期**：2016-08-16

基金项目：国家社会科学基金项目（14BYY096）；国家自然科学基金项目（61402419、61272221）；国家 863 高技术研究发展计划基金项目（2012AA011101）；计算语言学教育部重点实验室（北京大学）开放课题基金项目（201401）；国家 973 重点基础研究发展计划基金项目（2014CB340504）；河南省高等学校重点科研基金项目（15A520098）

作者简介：张金瑞（1990-），男，河南许昌人，硕士，研究方向为自然语言处理；柴玉梅（1964-），女，河南郑州人，教授，研究方向为机器学习、数据挖掘和自然语言处理；咎红英（1966-），女，河南焦作人，教授，研究方向为中文信息处理；高明磊（1963-），男，河南郑州人，硕士，研究方向为数据挖掘。E-mail：zjr__zhengda@139.com

主题与各个类别之间的距离，取其距离最小者作为该主题所属的类别。本文还保持了文档的词序，在此前提下，生成文档中的词语时，加入了二元语法[7]，当时是用来做短语发现和信息检索的，现在我们把它用于文本分类中。VB-LDA 是一种弱监督的学习方法，在数据集上的实验结果也验证了本文所提方法的有效性。

1　文本分类算法 VB-LDA

本文用 VB-LDA 算法对文档分类的主体思路如下：首先用 LDA 改进模型对文档集生成主题模型，然后获取主题的高频词和类别的代表词[8]，将它们都转化成词向量，最后用距离度量来计算出每篇文档中概率最大的主题所对应的类别，即为该文档的类别。文中需要用到的参数和其定义见表 1。

表 1　各个参数和其定义

参数	定义
T	主题的数目
D	文档集
n_d	文档 d 中词的个数
θ_d	文档 d 中的主题概率分布
z_i^d	文档 d 中第 i 个词所分配的主题
x_i^d	文档 d 中第（$i-1$）个词和第 i 个词之间的二元状态变量
w_i^d	文档中第 i 个词
Φ_z	主题 z 对没有形成二元语法的词的概率分布
φ_{zw}	词 w 选择主题 z 时，与下一个词之间的二元状态变量的分布
σ_{zw}	主题 z 对于形成二元语法的词的概率分布
α、β、γ、δ	分别为 θ、Φ、φ、σ 的先验参数

1.1　基于 LDA 改进的文档集的生成

LDA 是一种概率生成模型[9,10]，主要用于发现文档集中的潜在语义结构，它是由文档、主题和词语组成的三层贝叶斯生成模型。其核心思想是文档可以表示为一系列潜在主题的混合分布，其中每一个主题代表了文档集中全部词的概率分布，与潜在主题相关的词的概率分布较高。

在 LDA 模型中，并没有考虑词序问题，是典型的词袋模型。文档中的词语相互独立，当前词的主题概率分布既不依赖于前一个词的主题概率分布，也不影响后一个词的主题概率分布，每个单词的生成都是一个独立的过程。

在本文提出的改进算法 VB-LDA 中，将文档中的词序加以考虑，不再是单纯的词袋模型。本文在给文档中的词选择主题时，加入了二元语法。即在两个相邻词语之间引入一个状态随机变量 x，如果 $x=1$，则这两个词形成一个二元语法，下一个词的主题选择直接受到前一个词的主题分布的影响；如果 $x=0$，则二者形不成一个二元语法，下一个词的主题选择是一个独立过程，不受其它因素影响。它和 LDA 的不同之处在于：文档中每个词的产生过程并不是独立的，它可能依赖于前一个词的主题概率分布，也可能影响下一个词的生成过程。因此，一个词语的生成不仅仅受到主题概率分布的影响，也受到一个随机的贝努利分布的影响，这个分布决定当前词和前一个词是否形成一个二元语法。LDA 改进模型的文档生成过程如算法 1 所示。

算法 1：LDA 改进模型的文档生成过程

输出：文档集 D
(1) **for each** 文档 $d\in D$ **do**
(2)　抽取其 $\theta_d\sim$Dirichlet(α)；
(3)　**for each** $w_i\in d$ **do**
(4)　　抽取其 $z_i^d\sim$Dirichlet(θ_d)；
(5)　　抽取其 $x_i^d\sim$Bernoulli($\varphi_{z_{i-1}^d w_{i-1}^d}$)；
(6)　　**if** ($x_i^d==1$) **then**
(7)　　　抽取第 i 个词 $w_i^d\sim$Multinomial($\sigma_{z_i^d w_{i-1}^d}$)；
(8)　　　**if** ($x_i^d==0$) **then**
(9)　　　抽取第 i 个词 $w_i^d\sim$Multinomial($\Phi_{z_i^d}$)；
(10)　**end if**
(11)　**end for**
(12) **end for**

1.2　主题高频词和类别代表词的获取

在上述 LDA 改进模型中，可以观察到文档中的词语，但不知道它内部隐藏的主题概率分布和每个主题对于文档集所有词的概率分布。然而，想要准确的计算出这两个隐藏的概率分布是很困难的，在本文中采用 Gibbs Sampling[11] 算法来近似的估计它们。在 Gibbs Sampling 算法中，由联合概率 $P(w,z,x\mid\alpha,\beta,\gamma,\delta)$，可以利用贝叶斯变换公式求出条件概率 $P(z_i^d,x_i^d\mid z_{-i}^d,x_{-i}^d,w,\alpha,\beta,\gamma,\delta)$[7]，其中 z_{-i}^d 代表除去 w_i^d 以外的其它词的主题分配情况，x_{-i}^d 是 w_{i-1}^d 跟除去 w_i^d 之外的其它词能否形成二元语法的情况。根据文献 [11]，在 Gibbs Sampling 过程中，为单词更新主题的条件概率公式如下

$$P(z_i^d,x_i^d\mid z_{-i}^d,x_{-i}^d,w,\alpha,\beta,\gamma,\delta)\propto(\gamma_{w_i^d}+p_{z_{i-1}^d w_{i-1}^d x_i}-1)(\alpha_{z_i^d}+q_{dz_i^d}-1)\times\begin{cases}\dfrac{n_{z_i^d w_i^d}+\beta-1}{\sum_{v=1}^{W}(n_{z_i^d v}+\beta)-1} & if\ x_i^d=0\\ \dfrac{m_{z_i^d w_{i-1}^d w_i^d}+\delta-1}{\sum_{v=1}^{W}(m_{z_i^d w_{i-1}^d}+\delta)-1} & if\ x_i^d=1\end{cases}\tag{1}$$

式中：n_{zw} 表示词语 w 未能与相邻词语形成二元语法时，有多少次被分配到主题 z 上。m_{zwv} 表示当词 w 和词 v 形成二元语法时，有多少次词 v 作为第二个词被分配到主题 z 上。p_{zwk} 中的 k 有两种取值 0 和 1，当 $k=0$，p_{zw0} 表示当词 w 的主题为 z 时，词 w 和下一个词的二元状态变量有多少次等于 0；当 $k=1$ 时，p_{zw1} 表示当词 w 的主题为 z 时，词 w 和下一个词的二元状态变量有多少次等于 1。q_{dz} 表示在文档 d 中，有多少个词语被分配到主题 z 上。以上所述的计数变

量都可以在文档集中观察到，知道了它们的值，可以按如下几个公式来估计后验参数 θ、Φ、φ 和 σ

$$\theta_z^d = \frac{q_{dz} + \alpha}{\sum_{t=1}^{T}(q_{dt} + \alpha)} \tag{2}$$

$$\Phi_{zw} = \frac{n_{zw} + \beta}{\sum_{v=1}^{W}(n_{zv} + \beta)} \tag{3}$$

$$\varphi_{zwk} = \frac{\gamma + p_{zwk}}{\sum_{k=0}^{1}(\gamma + p_{zwk})} \tag{4}$$

$$\delta_{zwv} = \frac{\delta + m_{zwv}}{\sum_{v=1}^{W}(\delta + m_{zwv})} \tag{5}$$

获取主题高频词的算法过程如算法 2 所示。

算法 2：主题高频词的获取

输入：α、β、γ、δ，D，T 和 Gibbs Sampling 时的最大迭代次数 M

输出：θ、Φ、σ，主题的高频词

（1）把式（2）～式（5）中的计数变量都初始化为 0；

（2）初始化文档集，为每个文档中的每个词语随机分配一个主题；

（3）重新计算式（2）～式（5）中计数变量的值；

（4）**while**（iteration＜M）**do**

（5）　**for each** 文档 $d \in D$ **do**

（6）　　**for each** $w_i \in d$ **do**

（7）　　除去 w_i 和 z_i^d 对计数变量的影响；

（8）　　用式（4）计算 w_i 和 w_{i-1} 之间的二元状态变量 x_i；

（9）　　**if**（x_i＝＝1）**then**

（10）　　用式（1）给 w_i 选择一个新主题；

（11）　　更新 q_{dz}，p_{zw1}，m_{zwv}；

（12）　　**end if**

（13）　　**if**（x_i＝＝0）**then**

（14）　　用式（1）给 w_i 选择一个新主题；

（15）　　更新 q_{dz}，p_{zw0}，m_{zw}；

（16）　　**end if**

（17）　**end for**

（18）　用式（2）来更新文档 d 的主题概率分布 θ_z^d；

（19）　**end for**

（20）　用式（3）～式（5）来估计后验参数 Φ_{zw}，φ_{zwk} and δ_{zwv}；

（21）**end while**

（22）利用得到的 Φ_{zw} 和 δ_{zwv} 可获得每个主题的高频词。

类别的代表词可用如下方法获得：用算法 2 对某类别的数据进行处理时，把 T 设为 1，待算法收敛时，得到的此主题的高频词即为该类别的代表词。

1.3 基于距离度量的 VB-LDA 算法

得到每个主题的高频词后，由这些高频词可确定该主题所对应的类别。Hingmire 等[9]都是人为来指定每个主题属于哪个类别，但在本文中不再人为指定，因为在人为指定时可能会产生一定的偏差，导致结果不太理想，影响算法的准确性。本文引入了词向量化工具 word2vec[12]，它可以把词语转化为词向量，即把词语数字化，以便于计算词语之间的距离。

word2vec 使用的是 distributed representation 的词向量表示方式，它的基本思想是通过训练，把每个词映射为 N 维实数向量（N 一般为模型中的超参），通过计算词与词之间的距离，比如余弦相似度、欧式距离等来判断它们之间的语义相似度，它采用一个三层的神经网络结构：输入层、隐藏层和输出层。这个三层的神经网络本身是用来对语言模型进行建模，但同时也得到一种单词在向量空间上的表示，而这个副产品才是 word2vec 工具的真正目标。

本文首先用词向量化工具 word2vec 训练语料，得到除停用词以外所有词的词向量。然后利用 Gibbs Sampling 算法对文档集进行采样得到每个主题的高频词，对每个类别的数据采样得到每个类的代表词。最后在已得到的词向量库中找到主题的高频词和类别的代表词所分别对应的词向量，即分别把主题的高频词和类别的代表词向量化。

在利用距离度量计算每个主题和各个类别之间的距离时，先计算该主题的每个高频词和某个类别的每个代表词之间的平均距离，将该主题的各个高频词与此类别代表词的平均距离之和作为该主题与此类别的距离度量值。把主题 t 中第 i 个高频词的词向量记为 V_i^t，类别 c 中第 j 个代表词的词向量记为 V_j^c。假设语料中共有 K 个类别 $\{C_1, C_2, \cdots, C_K\}$，主题的高频词和类别的代表词都共有 N 个。本文假设用 d_{ij} 表示词向量 V_i^t 和词向量 V_j^c 之间的向量距离。假如 $V_i^t = \{x_1, x_2, \cdots, x_{50}\}$，$V_j^c = \{y_1, y_2, \cdots, y_{50}\}$，则 d_{ij} 的计算公式如下所示

$$d_{ij} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_{50} - y_{50})^2} \tag{6}$$

用 d_{ip} 表示主题 t 中第 i 个高频词和类别 c 中各个代表词之间的平均距离。则 d_{ip} 按如下公式计算

$$d_{ip} = \sum_{j=1}^{N} d_{ij} / N$$

用 d_{tc} 表示主题 t 与类别 c 的距离，则 d_{tc} 的计算公式如下

$$d_{tc} = \sum_{i=1}^{N} d_{ip}$$

用 C 代表类别集合，即 $C = \{C_1, C_2, \cdots, C_K\}$。

根据上文得到的每篇文档的主题概率分布 θ_z^d，可找出文档 d 中概率最大的主题 t，再将获取的主题 t 的高频词和各个类别的代表词分别向量化，此时基于距离度量的文本分类算法 VB-LDA 如算法 3 所示。

算法 3：基于距离度量的文本分类算法 VB-LDA

输入：文档 d 中概率最大的主题 t 的词向量（V_1^t，…V_i^t…，V_n^t）**和所有类别的词向量**（$V_1^{C_1}$，…$V_j^{C_1}$…，$V_n^{C_1}$），（$V_1^{C_2}$，…$V_j^{C_2}$…$V_n^{C_2}$），…，（$V_1^{C_K}$，…$V_j^{C_K}$…，$V_n^{C_K}$）

输出：文档 d 的类别 C_d

（1）distances=null；//**创建空数组** distances，**用来存放** d_{tc}

（2）**for each** $c \in C$；

（3） d_{tc}；

（4） **for each** $V_i^t \in \{V_1^t, \cdots V_i^t \cdots, V_n^t\}$

（5） sum=0，d_{tp}；

（6） **for each** $V_j^{C_k} \in \{V_1^{C_k}, \cdots V_j^{C_k} \cdots, V_n^{C_k}\}$

（7） **按式**（6）**计算** d_{ij}；

（8） sum=sum+d_{ij}；

（9） **end for**

（10） d_{tp}=sum/N；

（11） $d_{tc}=d_{tc}+d_{tp}$；

（12） **end for**

（13） **将** d_{tc} **加入数组** distances；

（14）**end for**

（15）**求出数组** distances **中的最小值，其下标所对应的类别即为** C_d

本文算法虽然不需要标注的训练数据，但是在分类过程中用到了词向量之间的距离计算，因此它是一种弱监督的分类算法。

2 实验与分析

为验证本文改进算法的有效性，这里将 VB-LDA **算法和** Classify-LDA[9] **算法以及现在广为应用的** SVM **算法（这里采用台湾大学林智仁等开发的** libsvm **工具包）在相同数据集上的运行结果作对比。然后与刘章等**[13] **提出的基于递归神经网络**[14] **的无监督模型** bRNN **方法做对比。用实验结果的宏平均** F **值的大小来衡量算法的性能。由于本文算法的后验参数是近似估计得来的，所以对于每组实验，重复做十次并计算其平均** F **值。**

2.1 数 据

用如下 3 **个数据集来衡量算法** VB-LDA，Classify-LDA **和** SVM **对数据分类的能力。**

（1）20Newsgroup：**此数据集共包含** 20 **个新闻组的数据，将它的“**bydate**”版本用来做实验，此版本的数据分为** 60% **的训练数据和** 40% **的测试数据，它共包括** 6 **大类。我们用训练数据构建分类器，然后用它去预测测试数据。**

（2）SRAA：Simulated/Real/Aviation/Auto UseNet data 2：**该数据集共包含** 73 218 **篇** UseNet **文章，它来源于** 4 **个讨论组：**simulated auto racing（sim _auto），simulated aviation（sim _aviation），real autos（real _auto）**和** real aviation（real _aviation）。**以下是它们以不同的组合方式来做实验的实验分组：**

1）sim _auto vs sim _aviation vs rea _lauto vs real _aviation

2）auto（sim _auto+real _auto）vs aviation（sim _aviation+real _aviation）

3）simulated（sim _auto+sim _aviation）vs real（real _auto+real _aviation）

这里把该数据集随机分成 80% **的训练数据和** 20% **的测试数据。**

（3）WebKB：WebKB **数据集共包含从大学计算机科学部门收集到的** 8145 **个网页数据。它们一共分为** 4 **类：**student，course，faculty **和** project。**此数据集被随机分为** 80% **的训练数据和** 20% **的测试数据。**

首先对数据进行预处理，除去文档里面的 HTML **标签和停用词。对于数据集** 20Newsgroups **和** WebKB，**主题的数目设为类别数的** 2 **倍。而对于数据集** SRAA，**在实验数据上学习** 8 **个主题，并且在** 3 **组实验中标记出这** 8 **个主题的类别。本文算法的先验参数** α、β、γ、δ **分别设为** 50/T，0.01，0.1 **和** 0.01。**主题的高频词数和类别的代表词数均设为** 20 **个。**

2.2 人工指定和距离度量的对比实验结果

此实验用到了 20Newsgroup **中** comp，politics **和** religion **这** 3 **个子数据集，表** 2 **和表** 3 **是在** comp vs politics **和** politics vs religion **实验中，在原** LDA **模型基础上，人工指定主题的类别和用距离度量来决定两种方法的结果对比。**

表 2 comp vs politics **中主题到类别的判断**

ID	主题中的高频词	Class（comp/politics）		实验结果 F 值	
		人工指定	距离度量	人工指定	距离度量
0	people know **president government** make work **fire fbi** state **weapons**	**politics**	**politics**		
1	**gun israel** people **law** file **guns government rights weapons against**	**politics**	**politics**	0.960	0.960
2	people **turkish armenian armenians** know turks **war** years **killed soldiers**	**politics**	**politics**		
3	**drivescsi card system** problem **bit** thanks **disk** know **drives**	**comp**	**comp**		

表 3 politics vs religion 中主题到类别的判断

ID	主题中的高频词	Class (politics/religion)		实验结果 F 值	
		人工指定	距离度量	人工指定	距离度量
0	people **gun** know **government president** file **guns fire** state **weapons**	**Politics**	**politics**		
1	people know make evidence argument well things question moral objective	Religion	politics	0.810	0.872
2	**god jesus christian bible** people **church christians** time life know	**Religion**	**religion**		
3	**israel turkish** people **armenian war** israeli **armenians government turks** turkey	**Politics**	**politics**		

从表 2 中可以看出，实验 comp vs politics 中的 4 个主题都比较容易判断它们的类别，两种方法判断的结果都一样，而且分类结果的 F 值（准确率和召回率的综合体现）也相同。而在表 3 中看到，实验 politics vs religion 中的主题 0、主题 2 和主题 3 在两种方法中它们的类别判断结果相同，而对于主题 1，人工指定的类别为 religion，用距离度量计算的结果为 politics。

在表 3 中，距离度量的分类结果的 F 值比人工指定的 F 值高 6.2%。这是因为对于实验 politics vs religion 中的主题 1，很难用人工来判断它属于哪个类，指定时难免会出现疏忽，导致实验结果不理想，而用本文提出的基于词向量的距离度量方法则可以准确的计算出主题 1 和类别 politics、religion 之间的向量距离，分别为 3.47 和 3.49，所以主题 1 的类别应该是 politics。

2.3 VB-LDA 和 Classify-LDA 以及 SVM 的实验结果对比

此实验用到了 3 个数据集 20Newsgroup，SRAA 和 WebKB，VB-LDA 和 Classify-LDA 以及 SVM 这 3 种算法在数据集上的实验结果对比见表 4。

表 4 在不同数据集上的实验结果

Dataset	# Topics	Text classification (Macro-F1)		
		ClassifyLDA	VB-LDA	SVM
20Newsgroups				
comp vs politics	4	0.962	0.985	0.983
religion vs sports	4	0.899	0.903	0.907
politics vs religion	4	0.875	0.889	0.891
comp vs religion vs sports	6	0.908	0.935	0.939
comp vs religion vs politics	6	0.884	0.929	0.944
comp vs religion vs sports vs politics	8	0.835	0.888	0.912
SRAA				
sim _auto vs sim _aviation vs real _auto vs real _aviation	8	0.747	0.768	0.813
auto vs aviation	8	0.918	0.930	0.933
simulated vs real	8	0.916	0.921	0.928
WebKB				
WebKB	8	0.653	0.687	0.725

从上表可以看出，VB-LDA 算法在 10 组对比实验上表现的几乎都要比 Classify-LDA 算法好，在 religion vs sports 和 simulated vs real 这两组实验上，两种算法表现出了几乎相同的分类能力。另外，在实验 comp vs religion vs sports 和 comp vs religion vs politics 中，本文算法 VB-LDA 在 F 值上相比较于 Classify-LDA 提高了 2.7%和 4.5%，而在实验 comp vs politics，religion vs sports 和 politics vs religion 中，F 值分别提高了 2.3%，0.4%和 1.4%，最后在 comp vs religion vs sports vs politics 这组实验中，F 值提高了 5.3%，由此可以看出，四分类实验中 F 值的提高比三分类实验的多，而三分类实验中 F 值的提高又比二分类实验的多，这表明本文算法 VB-LDA 在类别较多的情况下效果较好。从结果中还可以看出，在 10 组对比实验上，VB-LDA 算法的分类能力和有监督的 SVM 算法的分类能力都比较接近，说明本文提出的方法在不需要标注数据时也能取得不错的分类效果。

类别 comp，politics，religion 和 sports 之间的距离见表 5。

表 5 可以看出，类别 comp 和 politics 之间的距离明显比类别 religion 和 sports、politics 和 religion 之间的距离都大，所以类别 comp 和 politics 比较容易区分，同时在分类实验中，comp vs politics 的 F 值比另外两组 religion vs sports、

表 5 类别之间的距离

class	class	distance
comp	politics	5.72
religion	sports	4.83
politics	religion	4.38

politics vs religion 的 F 值都要高，也验证了这一点。从中还看到，类别 religion 和 sports 之间的距离比类别 politics 和 religion 之间的距离大，相应的，religion vs sports 的 F 值比 politics vs religion 的 F 值高一点。

2.4 VB-LDA 和 bRNN 的实验结果对比

此实验用到了 20Newsgroup 中 comp，politics，religion，sciences 和 sports 这 5 个子数据集。在用 bRNN 方法来分类时，参数设置如下：输入层数大小设为 300，后传步数设为 5，隐含层大小设为 200，输出层大小设为 300。VB-LDA 和 bRNN 两种方法在数据集上的结果见表 6。

表 6 VB-LDA 和 bRNN 在 20Newsgroup 上的实验结果

Data	# Topics	Text classification (Macro-F1)	
		bRNN	VB-LDA
comp vs religion	4	0.921	0.973
comp vs sciences	4	0.866	0.864
politics vs sciences	4	0.881	0.923
politics vs religion vs sports	6	0.869	0.909
comp vs religion vs sciences	6	0.831	0.852
comp vs religion vs sciences vs sports	8	0.815	0.847

通过对比，可以看到，除了在实验 comp vs sciences 中，VB-LDA 和 bRNN 两种方法的分类效果几乎相同外，在其它几组实验中，VB-LDA 具有比 bRNN 更显著的分类能力，这说明本文算法 VB-LDA 在面对其它无监督文本分类算法时，也能取得更好的效果。

3 结束语

本文提出了一种基于 LDA 的弱监督文本分类算法 VB-LDA，该算法在为主题选取类别时，不再需要人工指定，而是先用 word2vec 把主题的高频词和类别的代表词向量化，而后利用距离度量计算当前主题与各个类别之间的距离来确定它属于的类别，避免了人工指定时可能会出现的疏漏。同时本文在保持文档词序的前提下，加入了二元语法，提供了更多丰富的信息，分类效果会更加明显。在数据集上的实验结果也验证了 VB-LDA 算法的有效性，表明该算法在不需要标注数据时，也能取得与 SVM 算法相当，比 bRNN 模型更好的分类能力。本文在给文档中的词语更新主题时，没有考虑句子与句子之间、段落与段落之间存在的主题转移问题，下一步将对此类问题做一些更深层次的研究。

参考文献：

[1] Blei DM，Ng AY，Jordan MI. Latent dirichlet allocation [J]. Journal of Machine Learning Research，2003 (3)：993-1022.

[2] Lee S，Baker J，Song J，et al. An empirical comparison of four text mining methods [J]. Journal of Computer Information Systems，2010，51 (1)：1-10.

[3] Zhu J，Ahmed A，Xing EP. MedLDA：Maximum margin supervised topic models for regression and classification [J]. Journal of Machine Learning Research，2012，13 (4)：2237-2278.

[4] Arora S，Ge R，Moitra A. Learning topic models-going beyond SVD [C] //Proceedings of the IEEE 53rd Annual Symposium on Foundations of Computer Science，2012：1-10.

[5] Anandkumar A，Foster DP，Hsu D，et al. Two SVDs Suffice：Spectral decompositions for probabilistic topic modeling and latent Dirichlet allocation [J]. Corr，2012.

[6] Swapnil Hingmire，Sandeep Chougule，Girish K，et al. Document classification by topic labeling [C] //In SIGIR，2013：877-880.

[7] Wang X，McCallum A，Wei X. Topical N-Grams：Phrase and topic discovery，with an application to information retrieval [C] //In Proc of ICDM，2007：697-702.

[8] Swapnil Hingmire，Sutanu Chakraborti. Sprinkling topics for weakly supervised text classification [C] //In ACL，2014：55-60.

[9] Blei D，Carin L，Dunson D. Probabilistic topic models [J]. Communications of the ACM，2012：77-84.

[10] Frigyik BA，Kapila A，Gupta MR. Introduction to the dirichlet distribution and related processes [J]. Spokesman Review，2010.

[11] GriffithsTL，Steyvers M. Finding scientific topics [J]. Proceedings of the National Academy of Sciences，2004，101 (S1)：5228-35.

[12] Tomas Mikolov，Kai Chen，Greg Corrado，et al. Efficient estimation of word representations in vector space [S]. Arxiv. org，2013.

[13] LIU Zhang，CHEN Xiaoping. Using word clustering to improve recurrent neural network language model [J]. Application of Computer System，2014 (5)：101-106 (in Chinese). [刘章，陈小平. 联合无监督词聚类的递归神经网络语言模型 [J]. 计算机系统应用，2014 (5)：101-106.]

[14] Mikolov T，Karafiát M，Burget L，et al. Recurrent neural network based language model [C] //Interspeech，Conference of the International Speech Communication，2010：1045-1048.