计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190900012

基于 SL-LDA 的领域标签获取方法

王 胜[1] 张仰森[1,2] 张 雯[1] 蒋玉茹[1,2] 张 睿[1]
1 北京信息科技大学智能信息处理研究所 北京 100101
2 国家经济安全预警工程北京实验室 北京 100044
(1028742881@qq.com)

摘 要 科学技术的发展为文献及学者的管理提出了新的挑战,为解决海量科技文献及学者的自动管理,文中提出了一种基于SL-LDA的领域标签获取方法。在海量科技文献的基础上,分析科技文献数据的分布特点,通过引入科技文献的词频特征构建了SL-LDA主题模型,利用该主题模型对同一学者的科技文献进行"主题-短语"抽取,获得初始领域关键词。接着引入领域体系,对主题模型的抽取结果与体系标签进行向量表征,经过位置特征加权后使用相似度进行体系映射,最终获得学者的领域标签。实验结果表明,在同样的文献数据量下,SL-LDA模型与传统的LDA模型、基于统计的TFIDF算法和基于网络图的TextRank算法相比,最终获取的标签词效果更好,准确率更高,F1值也提升到0.572,说明基于SL-LDA的领域标签抽取方法在学术领域具有较好的适用性。

关键词:领域标签;SL-LDA模型;标签映射;主题短语抽取;科技文献
中图法分类号 TP391.1

Domain Label Acquisition Method Based on SL-LDA Model

WANG Sheng[1],ZHANG Yang-sen[1,2],ZHANG Wen[1],JIANG Yu-ru[1,2] and ZHANG Rui[1]
1 Institute of Intelligent Information Processing,Beijing Information Science and Technology University,Beijing 100101,China
2 Beijing Laboratory ofNational Economic Security Early Warning Engineering,Beijing100044,China

Abstract The development of science and technology poses new challenges for the management of literature and scholars. In order to solve the problem of automatic management of massive scientific literature and scholars,this paper proposes a domain label acquisition method based on SL-LDA. On the basis of massive scientific literature,the distribution characteristics of scientific literature data are analyzed,and the SL-LDA theme model is constructed by introducing the word frequency feature of scientific literature. The theme model is used to extract the "theme-phrase" from the scientific literature of the same scholar and get the initial domain keywords. Then the domain system is introduced,the extraction results of the theme model are vector-represented with the system label. After the position feature weighting,the similarity is used for system mapping. Finally,the domain label of the scholar is obtained. Experiment results show that,compared withthe traditional LDA model,the statistical-based TFIDF algorithm and the TextRank algorithm based on network graph,the final label words obtained by SL-LDA model have better effect and higher accuracy with the same amount of literature data,and the F1 value is also raised to 0.572,indicating that the domain label acquisition method based on SL-LDA has good applicability in the academic field.

Keywords Domain tags,SL-LDA model,Label mapping,theme phrase extraction,Scientific literature

1 引言

经济社会的蓬勃发展促使各种科技项目不断产生,项目从立项、评审到验收均需要前沿学者的参与。在以往的经验中,学者的遴选往往由专人进行,人为统计各学者的研究领域,选择与项目领域相符的学者。然而,传统方法往往有以下缺点:同一时间内存在大量项目需要前沿学者参与,这无形中加大了人为遴选的工作量;人为遴选容易受到人的主观性和

到稿日期:2019-08-31 返修日期:2019-11-04 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金项目(61772081,61602044);科技创新服务能力建设-科研基地建设-北京实验室—国家经济安全预警工程北京实验室项目(PXM2018_014224_000010)
This work was supported by the National Natural Science Foundation of China (61772081,61602044) and Construction of Technological Innovation Service Capability-Construction of Research Base-Beijing Laboratory-National Economic Security Early Warning Project Beijing Laboratory Project(PXM2018_014224_000010).
通信作者:张仰森(zhangyangsen@163.com)

局限性的影响，且在整个遴选过程中容易受到自身的知识层次、社会关系、个人偏好与利益等因素的影响，对学者的领域判断不全面，进而影响遴选结果的准确性。

鉴于以上原因，本文提出了基于主题模型的学者领域标签自动获取方法。在海量科研数据库与针对性领域体系的基础上，从大规模学术数据中自动获取学者的领域标签。领域标签的自动获取对将来学者的自动检索、自动推荐等需求具有重要意义。

2 相关工作

领域标签获取主要分为传统的领域标签获取和基于关键词的领域标签获取。

在传统方面，Budura 等[1]利用学者在互联网各平台的简介进行抽取，在此称其为网络标签。区别于抽取的标签，网络标签通常由本人或他人总结添加，且没有统一的规范，用词随意，进而导致获取的标签复杂多样，可用性低。另外，由于互联网内容较为随意，且多带有作者的写作特点，这导致在标签抽取过程中难以区分正确的数据信息和无用的信息，抽取时往往要针对特定平台、特定学者设计具有针对性的抽取方案，无形中增加了工作量。Khan 等[2]基于本体技术，设计了一种 P2P 模式的学者研究领域管理系统，利用 RDF 技术解决学者领域获取的问题，但由于该方法使用了特定的模板，导致方法的扩展性不足。Zhang 等[3]利用 J2EE 技术设计了一套学者信息管理系统，通过人工更新学者的基本信息和研究领域信息等，并提供学者咨询推荐模块，通过使用 Pearson 相似度来计算用户问题文本与学者研究领域的相似性以实现学者推荐功能，该方法需要学者登陆网站进行信息的更新与维护，只适应少量学者单一领域的情况，无法完成大规模学者发现与机器化领域标签抽取的功能。

基于关键词的领域标签获取有多种抽取方法，常用的关键词抽取基础包括统计、主题、网络图等。关键词是对文章具有高度概括性的一系列词语，自动关键词抽取是识别文本中具有代表性的词语或短语的技术。自该领域被提出以来，相关研究人员相继提出了各种各样的方法，总体上分为有监督和无监督两大类，其中有监督的方法需要人为标注语料，适用于小文本的情况，但随着海量互联网数据的增加，人工标注的成本越来越大，近年来逐步转向无监督的方法。统计法的核心思路是通过文本中词语的统计信息来进行抽取，该方法无需训练数据，直接使用词频、位置等进行判断筛选。例如 Dam 等[4]将 TFIDF 算法用于恶意行为提取，利用加权因子和术语权重修正 TFIDF 结果的权值，进而提升关键词的提取效果。Zhao 等[5]利用 N 元语言模型和文档权重归并实现学者领域的自动识别，N 元语言模型直接使用词序进行计算，无需对文档进行分词、特征提取等操作即可对文档进行领域分类，进而找出学者的领域标签。主题模型利用概率分布来实现关键词的抽取，目前主要流行的是 Blei 等[6]提出的 LDA（隐含迪利克雷分布）模型。Groof 等[7]为解决 LDA 实验中多词实验结果不一致的现象，在采样阶段使用变分式吉布斯采样替代塌陷式吉布斯采样，使主题识别的准确性得到了提升。Zhou 等[8]为了减少微博热点事件中数据稀疏的问题，将时间序列特征与词频加权特征引入 LDA 算法，该方法得到的话题关键词具有较高的可解释性，同时能够较好地表明话题所展示的内容。Hu 等[9]提出了一种基于 LDA 的热点检测方法，该方法首先构建了分布式 LDA 计算模型用于计算热点，随后使用 Jenson-Shannon 距离计算各主题之间的相似度，将其用于追踪主题趋势的演变趋势。在网络图方面，基于 PageRank 改编的 TextRank 算法[10]最为著名。Wen 等[11]基于 TextRank 的基本思想建立了候选关键词图模型，使用 Word2Vec 来计算关键词之间的相似度，将计算结果作为节点权重的转移概率，最终使 TextRank 算法获得的结果具有词相关性。Li 等[12]则首先构建关键词矩阵，在关键词矩阵的基础上使用 textrank 算法对短文本进行关键词抽取，提高了算法的精度。

本文使用基于主题的抽取方法，将科技文献集合看作待抽取语料，使用改进的 SL-LDA 主题模型对语料进行“主题-短语”抽取，得到主题分布矩阵，实现标签的自动获取。

3 基于 SL-LDA 的领域标签获取方法

3.1 概述

本文提出的基于主题模型的领域标签获取方法的框架图如图 1 所示。

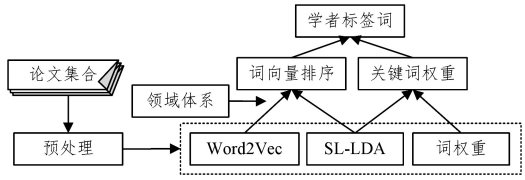

图 1 系统框架图

Fig. 1 System framework

该框架主要包含 4 个部分：1）数据预处理，获取初始数据集合；2）关键词抽取，通过 SL-LDA 进行“主题-短语”抽取，根据位置对短语进行权重赋值，并使用 word2vec 进行向量表征；3）领域体系映射，将“主题-短语”映射到体系，实现学者领域的统一管理；4）综合排序，对向量表征结果与权重赋值结果进行加权排序，通过阈值获得最能代表学者的标签词。

3.2 面向科技文献的 SL-LDA 模型

科技文献的更新速度快、结构固定，且通常具有较多的专业词汇，这对关键词的抽取带来了一定的挑战。部分文献虽已给出关键词，但仅由该部分数据还无法全面反映学者的学术领域方向。因此，本文引入主题的概念，从主题的角度对关键词进行抽取，使抽取出的关键词更贴近主题。

在基于主题的关键词抽取模型中，主流的是 LDA 主题模型，该模型认为一批文档中含有多个主题，而每个主题又可以用一系列短语来近似表示。一个文档的形成过程是，通过一定的概率选择一个主题，随后通过一定的概率选择当前主题下的短语，重复此过程直到形成一篇文档。而 LDA 主题抽取过程则是上述过程的逆操作，LDA 主题模型通常在新闻领域应用得较广[13]，然而在科技文献领域中，主题建模的效果会

受到科技文献特殊词频分布的影响。因此，本文对 LDA 模型进行改进，提出针对科技文献特殊词频分布的 SL-LDA 模型。

LDA 模型通过 Gibbs 抽样获得抽样参数 $\emptyset$ 和 θ，获取参数 $\emptyset$ 和 θ 的目的是构造一个收敛的马尔可夫链，进而从马尔可夫链中抽取合适的样本[14]。LDA 对短语的分配过程即为对 z_i 的抽样。其中，z_i 的后验公式如式(1)所示：

$$P(z_i=j|z_{-i},w)\rightarrow$$
$$P(w_i|z_i=j,z_{-i},w_{-i})P(z_i=j|z_{-i}) \quad (1)$$

已知 $P(w|z)$ 仅与 $\emptyset$ 相关，因此通过在 $\emptyset$ 上的积分，得到式(2)：

$$P(w_i|z_i=j,z_{-i},w_{-i})=\int P(w_i|z_i=j,\emptyset^{(j)})P(\emptyset^{(j)}|z_{-i},w_{-i})d\emptyset^{(j)} \quad (2)$$

其中，$\emptyset^{(j)}$ 是“主题-短语”的多项式分布，遵循式(3)：

$$P(\emptyset^{(j)}|z_{-i},w_{-i})\rightarrow P(w_{-i}|\emptyset^{(j)},z_{-i})P(\emptyset^{(j)}) \quad (3)$$

另外，$P(\emptyset^{(j)})\sim Dirichlet(\beta)$，同时，$P(\emptyset^{(j)})$ 是 $P(w_{-i}|\emptyset^{(j)},z_{-i})$ 的先验分布，因此对后验概率 $P(w_{-i}|\emptyset^{(j)},z_{-i})\sim Dirichlet(n_{-i,j}^{(w)}+\beta)$ 进行积分，即可获得式(4)：

$$P(w_i|z_i=j,z_{-i},w_{-i})=\frac{n_{-i,j}^{(w)}+\beta}{n_{-i,j}^{(*)}+V\beta} \quad (4)$$

同理可知，$P(z)$ 仅与 θ 有关，因此通过在 θ 上的积分可得式(5)：

$$P(z_i=j|z_{-i})=\frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,*}^{(d_i)}+T\alpha} \quad (5)$$

结合式(1)、式(4)、式(5)得到式(6)：

$$P(z_{-i}=j|z_{-i},w)\rightarrow\frac{n_{-i,j}^{(w)}+\beta}{n_{-i,j}^{(*)}+V\beta}*\frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,*}^{(d_i)}+T\alpha} \quad (6)$$

经过前文的计算，得到了 LDA 的非标准分布，但需要除以所有“主题-短语”分配的概率和，如式(7)所示：

$$P(z_i=j|z_{-i},w_i)=\frac{\frac{n_{-i,j}^{(w_i)}+\beta}{n_{-i,j}^{(*)}+V\beta}*\frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,*}^{(d_i)}+T\alpha}}{\sum_{j=1}^{T}\frac{n_{-i,j}^{(w_i)}+\beta}{n_{-i,j}^{(*)}+V\beta}*\frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,*}^{(d_i)}+T\alpha}} \quad (7)$$

其中，w_i 为第 i 个词语，$z_i=j$ 为将当前主题 j 分配给当前词 w_i，z_{-i} 为分配给非 z_i 的词权重和，$n_{-i,j}^{(w_i)}$ 表示主题为 j 且与词语 w_i 相同的权重和，$n_{-i,j}^{(d_i)}$ 表示文档 d_i 中主题为 i 的词语的权重和，$n_{-i,*}^{(d_i)}$ 表示当前文档中拥有主题的词语的权重和，V 表示词库大小，T 表示主题个数，$P(z_i=j|z_{-i},w_i)$ 为经过重新计算的后验概率。

另外，通过对学者科技文献的统计分析发现，科技文献的频次信息满足幂函数分布，图 2 给出了科技文献前 2000 个高频次词，其中横坐标表示高频词按照频次降序排序后对应的序号，纵坐标为高频词频次。

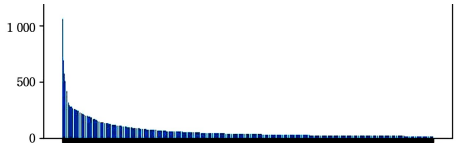

图 2　文档词频分布

Fig. 2　Document word-frequency distribution

经过统计可知，词频排名前 10% 的单词占据了全部科技文献词集的 81.1%，符合 Zipf 分布[15]，且通过对词频的研究发现，最能代表主题的词往往不是极高频词与极低频词，而是频次较靠前的中高频词。如果直接使用 LDA 模型对文档进行提取，则会造成某些中频词的缺失，同时用于词频较高的特征词通常结对出现，通常高频词被分配到主题中的概率较大，因此导致各话题的区分度不高，在前期的预处理中虽然进行了停用词过滤，但仍不能做到完全的过滤。

因此，本文提出了一种词频加权的 LDA 主题模型，首先统计文档中的词频信息，将词频特征引入 Gibbs 采样的过程中，降低高频词的影响力，提高中频特征词的影响力，接下来构建 SL-LDA 模型，使得模型不过分偏重于高频特征词词语。模型的词频加权公式如下：

$$C_i=1-\frac{|n_i-n_{mid}|}{max(n_{max}-n_{mid},n_{mid}-n_{min})}+1 \quad (8)$$

$$F_i=\frac{C_i\times\sum_{j=1}^{n}na_j}{\sum_{k=1}^{n}(na_k\times C_k)} \quad (9)$$

其中，n_i 表示当前词的词频；n_{mid} 表示选择中频词的词频；n_{max} 表示词频统计结果中的最大值；n_{min} 表示词频统计结果中的最小值；C_i 表示当前词的权重，取值范围为[1，2]。为保障加权后总特征词的个数不变，需要对每个特征词的权重做调整，其中，F_i 为特征词调整后的权重，$\sum_{j=1}^{n}na_j$ 为当前词出现的个数，$\sum_{k=1}^{n}(na_k\times C_k)$ 为所有词的权重和。用计算得到的 F_i 替换 Gibbs 采样过程中初始化的随机值。

3.3　word2vec 与学术词权重

Word2vec 是由 Google 公布的一款词向量计算工具[16]，利用深度学习技术在百万级词典与数亿级的训练语料上进行训练，训练得到的结果即为词向量模型，词向量有效地在空间中表达了词的语义信息。向量的训练模型指的是浅层神经网络 CBOW 或 Skip-gram 模型，其中 CBOW 模型如图 3 所示。

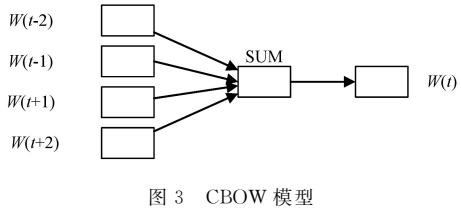

图 3　CBOW 模型

Fig. 3　CBOW model

CBOW 模型的特点是根据上下文来预测当前词，在训练时，首先为所有词语初始化一个 N 维词向量，并将窗口期内的上下文词进行累加，同时根据词频构建 Huffman 树来获得 Huffman 路径，根据路径计算叶节点的概率，随后采用梯度下降的方法调整非叶节点的参数和上下文的词向量，进行多次迭代后使结果收敛于真实结果。

学术论文通常分为标题、摘要、关键词、内容等信息。根据以往的经验，标题往往蕴含全文的中心思想，是全文内容的重要总结[17]，因此需要增加标题中词语的最终权重。关键词部分对全文主旨也具有一定的代表能力，而摘要部分则认为是对全文内容的简要概括，因此适当增加出现在

上述位置词语的权重。

3.4 领域词映射与标签词获取

针对各学者的科技文献，由于通过主题模型获得的短语存在差异性，无法对学者进行统一管理，因此引入学术领域体系来实现对学者的统一衡量。将主题模型的结果映射到领域体系中，映射公式如下：

$$sim(A,B)=\frac{\sum_{i=1}^{n}A_i\times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{10}$$

$$F_{(A,B)}=sim(A,B)*C_A*L_A \tag{11}$$

$$C_B=\begin{cases}C_B, & C_B>F_{(A,B)}\\ F_{(A,B)}, & C_B<F_{(A,B)}\end{cases} \tag{12}$$

其中，A 为主题模型获得的短语，B 为体系词，使用向量模型获得对应的词向量，对于未登录词则使用字向量拼接成词向量，$sim(A,B)$为最终计算的余弦相似度，C_A为主题模型分配的概率，L_A为短语在文档中的位置系数，$F_{(A,B)}$为经过加权得到的相似度，C_B为体系词的最终得分。经过上述计算后，将体系词经过排序后取 TOPN 即为映射结果。

4 实验设计与分析

为获取尽可能真实的实验数据，本文使用网络爬虫技术，爬取 CNKI 与万方数据库中的论文数据源，数据使用 jieba 进行分词处理，向量表征使用腾讯 AI 实验室公布的词向量模型[18]，本节将从评价标准介绍、数据预处理、SL-LDA 模型参数主题数的选择、基于主题模型的标签算法的评价 4 个方面来进行介绍。

4.1 评价标准介绍

由于 LDA 主题模型属于无监督模型，因此没有比较直观的评价标准来衡量模型的好坏。本文选择利用主题模型的“主题-短语”矩阵来进行评价，引入困惑度[19]作为模型的评价标准，通常认为困惑度越低，模型的效果就越好。困惑度的计算公式如式(13)和式(14)所示：

$$Perplexity(D)=exp\left(\frac{-\sum\log p(w)}{\sum_{d=1}^{M}N_d}\right) \tag{13}$$

$$p(w)=p(z|d)*p(w|z) \tag{14}$$

其中，$Perplexity(D)$表示当前模型的困惑度，d 表示科技文献文本，M 表示科技文献的个数，$\sum_{d=1}^{M}N_d$表示当前语料中所有词的数量和，$p(w)$为词 w 出现在矩阵中的概率，$p(z|d)$表示科技文献 d 为主题 z 的概率，$p(w|z)$表示词 w 出现在主题 z 中的概率。困惑度衡量主题模型预测出的结果与原样本信息的符合度[20]。

在计算标签准确度时，使用 F1 值来衡量学者标签的准确性，随机挑选多位学者，人工参照领域标签并结合对学者的了解来选择 4 个最合适的标签作为正确标签。根据算法排序得到的前 4 个标签与正确标签，使用上述指标进行评价计算，并最终计算平均 F1 值，计算公式如下：

$$P=\frac{\sum_{i=1}^{N}\frac{h_i\cap m_i}{h_i}}{N} \tag{15}$$

$$R=\frac{\sum_{i=1}^{N}\frac{h_i\cap m_i}{m_i}}{N} \tag{16}$$

$$F1=\frac{2*P*R}{P+R} \tag{17}$$

其中，h_i表示标准标签的个数，m_i表示算法得到的标签个数，$h_i\cap m_i$表示算法得到的正确的标签数，N 为样本总数。

4.2 数据预处理

为消除多数据源爬取造成的重复数据对计算结果的影响，在预处理阶段需要对数据进行去重处理，使用字共现、作者共现与关键词重合率来构建清洗模型。对于两个待比较的文本，首先判断其 DOI 是否相同，直接过滤掉具有相同 DOI 的文本，对于 DOI 不相同或不存在的文本，首先使用字共现对标题进行判断，若共现度大于 80% 则继续判断作者的共现次数和关键词共现次数，若作者的共现次数大于 1 且关键字共现率大于 0.5，则判定结果为重复并清除，共现公式如下：

$$coexist(A,B)=\frac{len(A\cap B)}{min\{len(A),len(B)\}} \tag{18}$$

在分词阶段，首先提取海量论文关键词数据，并将其添加到分词工具的用户词典，同时在计算之前使用 TextRank 抽取关键短语，将关键短语也添加到分词工具的用户词典中。另外，按照词频对整体语料数据排序，人工筛选高频无关词，将无关词添加到分词工具的停用词表中。

4.3 SL-LDA 模型参数的选择

主题模型中主题数的选取是影响主题聚类结果的重要因素。如果主题数设置得过小，则会导致模型聚类结果没有区分度；如果主题数设置得过大，则会导致将当前文档错误地划分到别的主题中。因此本文通过实验来计算不同主题数下的困惑度，并根据困惑度来确定最终的主题数。实验在其余参数不变的情况下只改变主题数，得到的实验结果如图 4 所示。

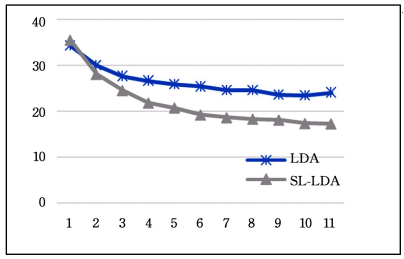

图 4 困惑度-主题数关系图

Fig. 4 Perplexity-topic num relation

图 4 中，LDA 曲线为 LDA 主题模型的困惑度曲线，SL-LDA 曲线为经过词频加权的 LDA 主题模型的困惑度曲线。关系图中，横坐标为主题数，纵坐标为困惑度。实验中，使用同一组参数来重复 3 次实验，取 3 次实验结果的平均值。

由实验结果可得，随着主题数的增加，3 种模型的困惑度都呈下降趋势，且都在主题数为 20 左右时下降幅度放缓甚至收敛，因此选取 20 为最优主题数。

4.4 标签算法的评价

引入领域体系来统一衡量学者的学术领域，领域体系参照国家自然科学基金领域体系制定，并在此基础上做适当修改，领域体系的示例如表 1 所列。

表 1　领域体系示例

Table 1　Example of domain system

一级	二级	三级	四级
计算机科学	计算机网络	网络安全	网络入侵检测
			隐私保护
		信息安全	信息隐藏
人工智能	自然语言处理	情感计算	舆情计算
		机器翻译	统计机器翻译
	机器学习	统计学习	

由于在学术领域标签抽取方面缺乏较权威的数据集，为验证算法的有效性，本次实验数据使用 12 位学者的学术论文数据，分别用 TFIDF 算法、TextRank 算法、LDA 算法和 SL-LDA 算法获取学者的学术标签并进行比较，具体示例如表 2 所列。在评价算法时，根据学者的主页介绍、招生简章等信息，人工选取体系中适当的词语作为学者领域标签的标准答案，将算法获取的标签词称作当前答案，将当前答案与标准答案进行效果评价。

表 2　算法抽取标签结果的对比

Table 2　Comparison of algorithm extraction label result

	Answer	SL-LDA	LDA	TFIDF	TEXTRANK
学者一	自然语言处理	机器学习	机器学习	机器学习	机器学习
	机器学习	自然语言处理	可视化分析	自然语言处理	自然语言处理
	知识工程	知识工程	知识库构建	可视化分析	可视化分析
	社会计算	知识库构建	络安全威胁	社会计算	识别模式
学者二	机器学习	机器学习	网络安全威胁	机器学习	网络安全威胁
	自然语言处理	可视化分析	可视化分析	可视化分析	可视化分析
	并行处理	自然语言处理	机器学习	知识库构建	机器学习
	识别模式	并行处理	识别模式	自然语言处理	识别模式

表 3　算法 F1 值的比较

Table 3　Comparison of algorithm's F1 value

Algorithm	4-2	4-3	4-4
TFIDF	0.444	0.381	0.375
TEXTRANK	0.389	0.405	0.354
LDA	0.361	0.357	0.313
SL-LDA	**0.572**	**0.476**	**0.458**

由表 2 可得，经过 SL-LDA 算法获得的标签与标准答案拥有较高的重合性，算法效果优于 LDA 与 TFIDF 算法。表 3 中的数据为利用式(15)－式(17)计算得到的 F1 值，其中 4-2 为当正确标签为 2 个时，与算法获得的 4 个标签进行计算获得的 F1 值；4-3 为当正确标签为 3 个时，与算法获得的 4 个标签进行计算获得的 F1 值；4-4 为正确标签为 4 个时，与算法获得的 4 个标签进行计算获得的 F1 值。经分析可得，在多种预测情况下，SL-LDA 的 F1 值高于传统 LDA 算法、基于统计的 TFIDF 算法和基于网络图的 TextRank 算法。这说明引入多词频特征加权的 SL-LDA 模型不仅能够从篇章分析文档的内容及其之间的联系，而且能够将学术数据降维，更适用于处理一定数量级的科技文献，有利于后续的标签映射与计算。该模型在一定程度上能够反映学者的研究方向，使用户能够较方便地对学者做全面了解，从而节省了用户的时间和精力，也间接说明了经过多词频特征加权的 SL-LDA 算法相较于传统算法能够更好地提取学术文本中的关键信息。

结束语　本文为实现学者研究领域的自动获取，提出了一种基于主题模型的领域标签获取方法。该方法对科技文献进行预处理后，获取文档的词频特征，并基于词频特征构建 SL-LDA 主题模型，对文档进行“主题-短语”抽取。其次，引入领域体系，将领域标签与主题模型获取的结果进行向量表征，将二者经过位置加权后的计算相似度进行体系映射，实现学者研究领域的获取。由实验结果可知，本文提出的方法较传统的 LDA 主题模型方法和 TFIDF 方法具有更好的效果。但在实际应用中，需要注意以下问题：1)学者的学术数据可获取，因为模型方法是建立在全文内容的基础上的；2)需要注意语料文档的类别，由于某些学者从事行政岗位后，发表的科技文献偏向教学培养等方向，与科研内容相差甚远，因此不能将模型运用于该方面的文档语料。在以后的工作中，将尝试对大规模学术数据的处理进行改进，缩短处理时间，优化处理逻辑和科技文献获取与筛选的方法，并继续完善 SL-LDA 算法，提升标签获取的准确率与效率。

参 考 文 献

[1] BUDURA A，BOURGES-WALDEGG D，R IORDAN J. Deriving Expertise Profiles from Tags[C]// Proceedings of the 2009 International Conferenceon Computational Science and Engineering. 2009：34-41.

[2] KHAN S，NABEEL S M. OPEMS：Online Peer-to-Peer Expertise Matching System[C]// Proceedings of the 1st International Conferenceon Information and Communication Technologies. 2005.

[3] ZHANG J. The design and implementation of expert information management system for think tank [D]. Harbin Institute of Technology，2017.

[4] DAM K H T，TOUILI T. Automatic extraction of malicious behaviors[C]// 2016 11th International Conference on Malicious and Unwanted Software (MALWARE). IEEE，2016.

[5] ZHAO H B，LU W. The Study of Expert Research Field Automatic Recognition [J]. New Technology of Library and Information Service，2010(2)：63-67.

[6] BLEI D M，NG A Y，JORDAN M I. Latent dirichlet allocation [J]. J Machine Learning Research Archive，2003，3：993-1022.

[7] GROOF R D，XU H. Automatic topic discovery of online hospital reviews using an improved LDA with Variational Gibbs Sampling[C]// IEEE International Conference on Big Data. IEEE，2018.

[8] ZHOU W X，ZHANG Y S，ZHANG L. Research on topic detection and expression method for Weibo hot events[J/OL]. Application Research of Computers. [2019-02-27]. https://doi. org/

10.19734/j.issn.1001-3695.2018.08.0601.

[9] HU X. News hotspots detection and tracking based on LDA topic model[C]// International Conference on Progress in Informatics & Computing. IEEE, 2017.

[10] MIHALCEA R, TARAU P. Textrank: Bringing order into text [C]// Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004: 404-411.

[11] WEN Y, YUAN H, ZHANG P. Research on keyword extraction based on Word2Vec weighted TextRank[C]// 2016 2nd IEEE International Conference on Computer and Communications (ICCC). IEEE, 2016.

[12] LI W, ZHAO J. TextRank Algorithm by Exploiting Wikipedia for Short Text Keywords Extraction[C]// International Conference on Information Science & Control Engineering. IEEE, 2016.

[13] CUI L, FAN M, YONG S, et al. A Hierarchy Method Based on LDA and SVM for News Classification[C]// IEEE International Conference on Data Mining Workshop. 2015.

[14] YANG C Y, PAN Y N, ZHAO L. Study on Topic Extraction of Literatures Based on Weighted Semantic and Citation Relation [J]. Library and Information Service, 2016, 60(9): 131-138, 146.

[15] CHEN Z, JI W. Exploiting noisy web data by OOV ranking for low-resource keyword search[C]// International Symposium on Chinese Spoken Language Processing. 2017.

[16] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv: 1301.3781, 2013.

[17] AO F, WANG L, CHEN M, et al. Text and position ranking algorithm based on sample weighted[C]// International Conference on Information Science & Engineering. IEEE, 2010.

[18] SONG Y, SHI S, LI J, et al. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Short Papers). 2018: 175-180.

[19] WU H, YIN S F, MA Y X, et al. WI-LDA: Technical Topic Analysis in Patents [J]. Library and Information Service, 2018, 62(17): 68-74.

[20] SHAN B, LI F. A Survey of Topic Evolution Based on LDA [J]. Journal of Chinese Information Processing, 2010, 24(6): 43-49, 68.

WANG Sheng, born in 1996, postgraduate. His main research interests include natural language processing and machine learning.

ZHANG Yang-sen, born in 1962, postdoctoral, professor, Ph.D supervisor, is a member of China Computer Federation (CCF). His main research interests include natural language processing and artificial intelligence.