doi:10.3772/j.issn.2095-915x.2020.03.002

一种基于概率潜在语义分析的专利主题标引方法研究

开放科学（资源服务）标识码（OSID）

包翔 刘桂锋

江苏大学科技信息研究所 镇江 212013

摘要：为了准确稳定地对专利主题进行标引，本文提出了一种基于概率潜在语义分析的专利主题标引方法。首先建立由共同主题和特定主题所联合起来的混合模型；然后通过这两类主题相关性推断出训练集和测试集特定主题的映射关系；最后选择相似度最高的主题作为专利的主题完成标引。实验结果表明，该方法能较为准确、稳定地对未标记专利进行主题标引。本文尝试将概率潜在语义分析运用到专利文本的标引中，既是对专利标引自动化的一种积极尝试，也为深层次挖掘专利信息情报技术提供了一种新的思路。

关键词：专利；标引；概率潜在语义；主题

中图分类号：G35

A Patent Topic Indexing Method Based on Probabilistic Latent Semantic Analysis

BAO Xiang LIU Guifeng

Institute of Scientific and Technical Information of Jiangsu University, Zhenjiang 212013, China

Abstract: In order to index the patent topics correctly and stably. We propose a novel patent topic indexing method based on probabilistic latent semantic analysis. In this method, a joint model based on shared topics and specific topics is established, a mapping relation between source set and target set can be induced, and we can index the patents based on the mapping matrix. It is experimentally demonstrated that the proposed method can index patents' topics correctly and stably. This research tries

基金项目：江苏省高校哲学社会科学研究一般项目“主题模型在高校图书馆知识产权信息服务中的研究与实践”（2019SJA1870）；江苏省高校自然科学研究面上项目“基于多示例多标签学习及深度神经网络的专利主题分类研究”（19KJB520005）。

作者简介：包翔（1991-），硕士，馆员，研究方向：机器学习、文本挖掘，E-mail: bx425bob@163.com；刘桂锋（1980-），博士，研究馆员，研究方向：数据治理、知识服务。

doi:10.3772/j.issn.2095-915x.2020.03.002

to apply probabilistic latent semantic analysis into patent indexing, which is not only a positive attempt to automate patent classification, but also provides a new perspective for deep mining patent information.

Keywords: Patent; indexing; probabilistic latent semantic analysis; topic

引言

在“知识产权强国”背景下，专利文献作为重要的知识产权信息，因其技术信息内容涉及人类生活的诸多方面，可以从侧面反映一个国家的创新能力、科技水平和市场化程度，是衡量科技产出和知识创新的一项重要指标[1]。因此，专利的分析与挖掘对于实现“知识产权强国”的目标显得非常关键。

专利标引，作为专利处理与挖掘的重要步骤，是指对一段专利文本所表达的中心思想或中心内涵进行提炼的过程。目的是为了反映出一段完整内容的基本思想，方便社会公众快速准确地从知识库中检索到自己需要的内容，对专利信息的情报挖掘起着基础性的作用。专利标引作为专利分析和挖掘的一个重要基石，越来越受到学界和业界的广泛关注。

专利的标引可以分为人工标引和自动标引，现在随着专利数据量的不断增加，对于专利自动标引方法的需求已经越来越迫切。自动标引大体可分为统计分析方法、语义分析方法、人工智能分析方法。统计分析方法主要利用专利信息中术语的显著特征，如共现、逆文档频次、互信息等。Wartena 等人[2]通过定义单词的共现分布，实现关键词的标引，提升了标引的效率；罗准辰等人[3]提出了以互信息与构造词串边界参数表的方法识别词串的方法，对于单独的关键单词提取和关键词串提取效果明显，提高了专利标引的精度与广度；李军锋等人[4]采用 K- 最邻近耦合图将专利文献映射成复杂网络图模型，分析关键词位置信息、关键词跨度信息以及关键词逆文档频率信息，完成专利关键词标引工作。语义分析方法则从自然语言的语义角度探索关键词标引，索红光等人[5]通过词汇间语义信息提出构建词汇链的算法，提高了标引精度；Noh 等人[6]利用主题相关度，通过分析候选关键词的语义信息从句子抽取关键词；丁杰等人[7]介绍了边界标记集的概念，并结合专利文献中术语边界的特点构建专利术语边界标记集，提出了一种种子术语权重计算方法抽取种子术语，实现对专利术语的标记；刘小蝶等人[8]归纳了专利文本中并列结构的语义、结构和外部词 3 个方面的特点，提出一种基于边界感知原则的识别方法，在概念层次网络 (HNC) 理论的基础上，从 8 个维度对并列结构进行标注，考察并总结语义特征、结构特征和外部词特征。人工智能的方法则主要是将机器学习模型与专利中文语义标注相结合[9-10]，以提高标引的准确率与召回率。

概率潜在语义分析 (Probabilistic latent semantic analysis，PLSA) 是由 Hofmann[11] 于 1999 年提出，其核心思想是使用概率统计模型来模拟文本中词的生成过程，该模型有更为合理的概率解释，而且进一步解决了同义词、多义词

的问题。在文档层面上，PLSA 可以将文档映射到各个主题，这些主题可以看作文本类别，每一文本所属类别中概率最大的那一类可作为文本最终所属类别。PLSA 在文本分析中有着广泛的应用，蒋铭初等人 [12] 将训练样本通过 PLSA 模型映射到隐语义空间，以文本的主题分布表示一篇文本，利用多标记假设重用算法进行文本分类；何炎祥等人 [13] 通过启发式初始化的 PLSA 模型训练得到贴近兴趣类别的主题模型，然后从训练结果中抽取可靠的话题并以此构建分类器，对用户的分享数据进行分类；吉余岗等人 [14] 提出了融合异质信息网络和主题模型构建方面分预测算法 (HINToAsp)，从意见短语角度构建了评论主题挖掘模型 (Phrase-PLSA)，有效整合评论信息和评分信息进行方面主题挖掘，进而提出了在“用户评论商品”异质信息网络上的主题传播模型。

PLSA 方法可以达到在维数降低的同时也保证了原本的语义空间结构的效果，而且面对专利文献晦涩难懂，专业词汇多的特点，PLSA 可以有效降低专利文本的维度，更好提取文本的语义进行分析。本文拟将 PLSA 方法应用于专利标引中，首先对已经标引过的专利语料库 (也称训练集) 和未标引的专利语料库 (也称测试集) 进行模型构建，之后利用两者之间主题的关系建立已标引专利数据与未被标引数据的映射关系，从而完成对未标引专利数据的标记工作。

1 相关理论

本文基于概率潜在语义分析 (PLSA) 模型，并且基于该模型挖掘出训练集和测试集的共同主题和特定主题的方法。PLSA 广泛应用于主题建模、信息检索、过滤、自然语言处理等领域，PLSA 假设一篇文档是由多个主题混合而成的，因而考虑到词分布和主题分布，最后使用期望最大值 (EM) 算法来学习参数。如图 1 所示，它通过单词 / 文档共现矩阵，在模拟文档的生成过程中涉及选择主题，然后从主题中选择单词的过程。具体过程包括：选择一篇文档，其被选中的概率为 $P(D=d_i)$；某个文档属于某一个主题的概率为 $P(Z=z_K|D=d_i)$；某个主题中某个词出现的概率为 $P(W=w_j|Z=z_K)$，$P(D=d_i)$，$P(Z=z_K|D=d_i)$，$P(W=w_j|Z=z_K)$ 是通过 EM 算法进行估计的。

在本文提出的方法中，假设专利的主题是由共同主题和特定主题共同表示的，特定主题包含训练集中的特定主题和测试集中的特定主题，一般来说，共同主题和特定主题的个数是根据数据集的情况而定义。

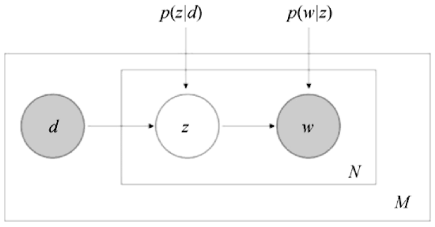


图 1 PLSA 示意图

如何计算训练集和测试集中主题之间相关性的计算方法是本文的一个关键问题。本文将共同主题作为一个桥梁，以希望找到训练集与测试集特定主题之间的关系。实际中，若训练集和测试集的主题都与共同主题有关联，它们很可能是语义相关的。

挖掘共同主题和特定主题，本方法借助一

doi:10.3772/j.issn.2095-915x.2020.03.002

个隐含的变量 Π∈[0,1]，来判断主题词是否来自于共同主题还是特定主题。最终通过以下公式 (1) 得到单词 / 文档共现矩阵 (d_i^l,w_j)，然后使用 PLSA 方法得到对应的主题和主题词。

$$P(w_j|d_i^l)=P(\prod=0|\mu_{d_i^l})\sum_{k=1}^{K}P(z_K|d_i^l,\prod=0)$$
$$P(w_j|z_K)+P(\prod=1|\mu_{d_i^l})\sum_{k=1}^{K}P(z_r^l|d_i^l,\prod=1)P(w_j|z_r^l) \quad (1)$$

而下一步重点就在于两个方面：其一就在于计算共同主题和特定主题之间的相似性；其二则是推断出训练集和测试集中特定主题之间的相关性。

为了解决第一个问题，本文采取 JS 散度 (Jensen-Shannon divergence) 作为文档 - 主题分布之间的相似度度量。JS 散度被广泛应用在度量概率分布之间的相似度。对于两个分布 P 和 Q，JS 散度表示为：

$$JSD(P\|Q)=\frac{1}{2}KL(P\|M)+\frac{1}{2}KL(P\|Q) \quad (2)$$

其中 $M=\frac{1}{2}(P+Q)$，$KL(.\|.)$ 是 KL 散度 (Kull-back-Leibler divergence)，它是衡量两个分布之间的相似度。所以共同主题 z_K 和特定主题 z_l^r 之间的相似度 $\theta_{z_Kz_r^l}$ 可以表示为：

$$\theta_{z_Kz_r^l}=JSD(P(D^l|z=z_K)\|P(D^l|z=z_r^l)) \quad (3)$$

即如果一个共同主题和特定主题总是在文档中同时出现，则它们之间有很小的 JS 散度，而这两个主题之间相似度会很大。

对于第二个问题，是推断计算出训练集和测试集中特定主题之间的相关性，但是由于训练集和测试集中特定主题不可能共现，所以不能用公式 (3) 来直接计算相似度。因此，本文选择皮尔森相关系数 (PCCs) 来计算他们的相似性，且当 PCCs 的绝对值越大时，两个变量之间的相关性越大，越接近 -1 表示负相关，相反，越接近于 1 表示正相关。本文中，训练集和测试集各自的特定主题之间的相关性可以表示如下：

$$\rho(z_i^t,z_j^s)=\frac{\Sigma_k(\theta_{z_K,z_i^t}-\bar{\theta}_{z_i^t})(\theta_{z_K,z_i^s}-\bar{\theta}_{z_j^s})}{\sqrt{\Sigma_l(\theta_{z_K,z_i^t}-\bar{\theta}_{z_i^t})^2}\sqrt{\Sigma_l(\theta_{z_K,z_j^s}-\bar{\theta}_{z_j^s})^2}} \quad (4)$$

其中，$\bar{\theta}_{z_i^l}$ 代表 z_i^l 和所有的共同主题之间的平均相似度，基于公式 (3)，可以得到训练集特定主题与测试集特定主题之间的映射矩阵 $U\in R^{K^t\times K^s}$，如下所示：

$$U=\begin{bmatrix}\rho(z_1^t,z_1^s) & \rho(z_1^t,z_2^s) & \cdots & \rho(z_1^t,z_{K^s}^s)\\ \rho(z_2^t,z_1^s) & \rho(z_2^t,z_2^s) & & \rho(z_2^t,z_{K^s}^s)\\ & \vdots & \ddots & \vdots\\ \rho(z_{K^t}^t,z_1^s) & \rho(z_{K^t}^t,z_2^s) & \cdots & \rho(z_{K^t}^t,z_{K^s}^s)\end{bmatrix} \quad (5)$$

通过映射矩阵 U，可以判断出训练集特定主题与测试集特定主题之间的相似程度，并且可以根据矩阵中的数值，判断出与测试集中主题最相关的训练集主题，从而实现自动的专利主题标引。

2 实验及分析

2.1 实验准备

专利 IPC 分类号是目前国际通用的专利文献分类和检索工具，它对于海量专利文献的组织、管理和检索，做出了不可磨灭的贡献。IPC 分类号一般采用等级形式，即将技术内容注明：部—大类—小类—大组—小组，是一种逐级分类而形成的完整的分类体系。

本实验在上海知识产权公共服务平台的中国专利数据库中选取水处理技术领域的 1000 篇专利文献作为语料库 [15]。实验选择 IPC 号为

D06(织物等的处理；洗涤；其他类不包括的柔性材料) 和 E03(给水；排水) 的部分专利文本作为实验数据。这些专利的 IPC 分类号中大类为 D06、E03 的各有 250 篇，具体来说，小类号为 D06M 有 78 篇，D06B 有 26 篇，D06F 有 86 篇，D06P 有 27 篇，E03C 有 56 篇，E03B 有 55 篇，E03D 有 84 篇，E03F 有 55 篇。为了更好地展示实验步骤，将训练集和测试集的个数调整到基本相同，如表 1 所示，将 IPC 分类号为 D06B、D06M、E03C 和 E03D 看作训练集，D06F、D06P、E03D 和 E03F 作为测试集，通过训练集中已标引的专利数据，建立标引模型而确定测试集中未标引数据对应 IPC 分类号的大类信息，以此实现专利的自动标引。

表 **1** 实验数据来源

训练集（已标注专利）	测试集（未标注专利）
D06B(纺织材料的液相、气相或蒸汽处理)	D06F(纺织品的洗涤、干燥、熨烫、压平或打折)
D06M(对纤维、纱、线、织物、羽毛或由这些材料制成的纤维制品进行D06类内其他类目所不包括的处理)	D06P(纺织品的染色或印花；皮革、毛皮或各种形状的固体高分子物质的染色)
E03C(干净水或废水的户内卫生管道装置)	E03D(冲水厕所或带有冲洗设备的小便池；其冲洗阀门)
E03B(取水、集水或配水的装置或方法)	E03F(下水道，污水井)

实验预处理的步骤主要包括分词、去噪和特征选择。分词步骤采用 jieba 中文分词的 .NET 版本并通过精确分词模式来实现，去噪步骤则是去除一些标点符号、常用词，例如“和、或、且、与、涉及、包括、根据、的”等词语。此外还考虑到专利文本的特殊格式，例如专利文本会在权利要求书中写出类似于以下的格式：“一种…方法 / 仪器 / 设备 / 工艺，其特征在于：…”。针对专利文本的特殊形式，本文将“一种、方法、仪器、设备、工艺、其、特征、在于、本发明”等词语删除，最终采用 TF-IDF 特征对文本的特征进行描述。经过以上的预处理步骤，本方法具体的分词结果示例如图 2 所示。

智能 控制型 抗冲击 城市 下凹式 雨水 公园 下凹式 雨水 公园 雨水 切换井 下凹式 雨水 公园 内 设有 地下水 池 地面 控制室 若干块 透水 地面 若干块 下凹式 绿地 若干个 渗透 井 若干个 储存 渗透池 渗透 井 通过 渗透 排水管 储存 渗透池 地下水 池 连通 雨水 切换井 与 市政 雨水 干管 连通 地下水 池 一路 通过 重力 式 进出水管 进 出水阀 和 排水泵 排水阀 连通 雨水 切换井 地面 控制 室内 设有 控制器 由 控制器 切换井 液位计 水池 液位计 送来 水位 信号 值 控制 进 出水 控制阀 排水阀 开通 关断 形成 智能 控制型 雨水 调蓄 缓排 结构 下凹式 雨水 公园 雨水 调蓄 缓排 智能化 管理 提高 城市 雨水 抗冲击 能力 充分发挥 雨水 公园 所 具有 娱乐 休闲 场所 功能

图 **2** 本文方法分词结果的示例

在特征选择阶段，文本特征利用 TFC 加权法 [16] 计算出每一个特征词的权重，本实验选择该方法是因为标题和摘要篇幅长度相差较大，而 TFC 加权法能消除篇幅长度对特征词权重的影响，并且选取了前 1000 个 TF*IDF 值对应的特征词作为数据的索引词。

doi:10.3772/j.issn.2095-915x.2020.03.002

2.2 实验结果与分析

按照本文第一部分的步骤进行模型构建，并在水处理数据集上进行验证，实验结果包括共同主题对应的主题词、训练集中特定主题对应的主题词、测试集中特定主题对应的主题词，以及映射矩阵。其中主题对应的主题词选取前十个密切相关的主题词，根据数据集的实际情况，训练集和测试集的 IPC 小类的个数都是 4。因此将训练集特定主题的个数和测试集特定主题的个数设为 4，并且也将共同主题的个数设为 4，具体的结果与分析如表 2 所示。

表 2 主题模型分析结果

共同主题	主题1	烘干 部件 隔板 凹槽 本体 支撑 横管 风道 固定 压力
	主题2	织物 染色 整理 材料 面料 纤维 加工 溶液 涂层 印花
	主题3	马桶 冲水 通道 冲洗 水箱 出水口 下水管 壳体 水路 排污
	主题4	雨水 机构 固定 污水 城市 卫生设备 出水管 雨水井 排水 升降杆
训练集特定主题	主题1	纤维 改性 表面 质量 纱线 纤维素 处理剂 棉纤维 抗菌 反应
	主题2	微胶囊 整理剂 供水管 原料 管网 涂敷辊 染色机 法兰 浆液 冷却
	主题3	取水 防臭 空气 输送 供水系统 水泵 箱体 水槽 出水口 水流
	主题4	收集 涤纶织物 雨水 蓄水箱 管道 盖板 主槽 储水容器 碳纤维 水井
测试集特定主题	主题1	衣物 滚筒 旋转 检查井 洗涤水 清洗 井筒 桶 元件 喷嘴
	主题2	染色 发酵仓 异味 纤维 控制系统 电机 部分 沉淀腔 流出物 疏通
	主题3	冲洗 冲水 便器 进水 水箱 开关 洗涤剂 区域 按键 真空
	主题4	组件 轴承 雨水井 明沟 布草 脱水蓝 模块 坐便器 阻尼器导槽 支承

本实验将共同主题的个数设置与训练集、测试集中的特定主题个数相同，并且将共同主题中每个主题对应的前十个词列出，表 2 中的共同主题的主题词都是整个是水处理领域中比较常用的词语，而且它可能与训练集和测试集的主题都相关。如表 2 所示，可以发现，共同主题中主题 1、主题 2 的对应的词语主要是 IPC 分类号为 D06(织物等的处理；洗涤；其他类不包括的柔性材料) 的相关技术词语，例如：织物、烘干、染色、面料、纤维、溶液、印花等词语。而主题 3 和 4 对应的词语主要是 IPC 分类号为 E03(给水；排水)，可以发现：马桶、冲水、水箱、出水口、污水等词语都与 E03 这个主题非常吻合。

从训练集中特定主题对应的词可以看出，主题 1 对应的词语与 IPC 分类号为 D06M (对纤维、纱、线、织物、羽毛或由这些材料制成的纤维制品进行 D06 类内其他类目所不包括的处理) 的专利语料库非常相关。主题 2 中的供水管、染色机、浆液、冷却等词会在 IPC 分类号 D06B(纺织材料的液相、气相或蒸汽处理) 的专利对应的技术中经常出现。主题 3 中的取水、防臭等词语是 IPC 为 E03C(干净水或废水的户内卫生管道装置) 的专利中经常会用到说法，而雨水、蓄水箱、储水容器、水井则与 E03B(取水、集水或配水的装置或方法) 密切

相关。

而测试集中特定主题对应的词也是与相对应 IPC 号所表示的内容非常相近的，主题 1 中的词大多都与纺织物的洗涤等操作相关，对应 IPC 为 D06F 的专利。主题 2 中主要关于纺织物的染色等操作，对应 IPC 为 D06P 的专利。主题 3 对应的词语能辨别出小便池设备，对应 E03D。主题 4 对应的词则是与下水道，污水井等词相关联，对应 E03F。并且，依据以上的分析，根据公式 (2-5) 可得本实验中的映射矩阵 U：

$$\begin{bmatrix} +0.156 & -0.786 & -0.125 & -0.875 \\ +0.114 & +0.834 & -0.394 & -0.917 \\ -0.265 & -0.642 & -0.747 & +0.920 \\ -0.217 & -0.762 & +0.506 & +0.956 \end{bmatrix}$$

从上述映射矩阵可以以下判断：本实验中，训练集、测试集中的特定主题个数都设置为 4，因此两者的相关矩阵为 4*4 的矩阵。矩阵的第一列代表的是测试集中的特定主题 1 与训练集中特定主题的相似程度，并以此类推。从上述矩阵中可以发现，测试集中的主题 1 与训练集中的主题 1、2 是正相关的，而与训练集中的主题 3、4 是负相关的，这也与实际情况相吻合，测试集中主题 1 的 IPC 号对应的是 D06F，而训练集中的主题 1、2 分别对应 D06B 和 D06M，训练集的主题 3、4 对应 E03C 和 E03B，D06M、D06F、D06P 都是 D06 大类下的小类。因此测试集中主题 1 与训练集主题 1、2 的相似性比较高，且为正数，而与主题 3、4 的相似度为负数。选取相似度最高的专利数据对应的 IPC 大类号就可以作为测试集的专利数据进行标引，如对于这个专利，可将其 IPC 的大类号确定为 D06。又例如，训练集中的主题 4 与测试集中的主题 4 的具有很多相同意义的词。例如雨水、水井等词语，同时也发现相关矩阵中两者的相似性最高，且达到了 0.956，很显然的就可以为测试集对应的专利进行标引，其 IPC 大类号与训练集的主题 4 对应的 IPC 大类号一致，可标引为 E03。综上所述，通过映射矩阵 U 可以得到对专利进行较为准确的主题标引。从而推测未知专利的 IPC 分类号实现标引。

由于本文提出的方法有随机初始化过程，因此在实验设计中，对于实验数据，运行程序 10 次，并对 10 次运行的平均结果进行分析。对比其他两种基于主题模型的标引方法，分别是基于统计生成模型的协同 - 对偶 PLSA 模型 [17]，以下简称 CDPLSA，和在 PLSA 模型的基础上运用非负矩阵三因子分解的主题标引方法 [18]，简称 TLPLSA 方法，并采用 P(Precision 准确率)、R(Recall 召回率)、F 值作为评价方法的指标。具体实验数据如表 3。

表 **3** 各种标注算法在数据集上的标引结果

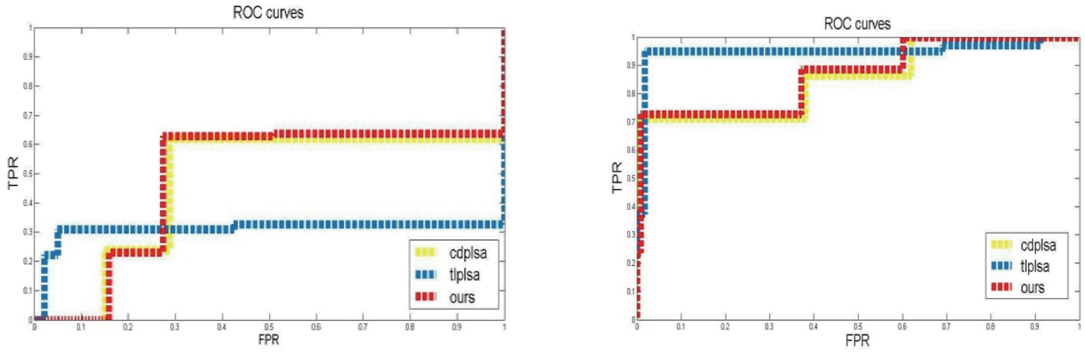
	D06 & E03		
	P	R	F
CDPLSA	0.667	0.666	0.666
TLPLSA	0.833	0.310	0.452
Ours	0.679	0.677	0.678

表 3 比较了三种方法在数据集上的标引指标，实验结果表明，该方法在数据集中 F 值都是最高的。TLPLSA 的方法在数据集上取得最优的 P 值，但是 TLPLSA 的 R、F 指标都是非常不理想的。综上所述，本文的方法与 CDPLSA 和 TLPLSA 这两种主题模型的衍生标注算法相比，本文的方法并未出现极端的运算结果，方法总体的稳定性较好，并且本文在中文专利

doi:10.3772/j.issn.2095-915x.2020.03.002

数据集的不同数据集上的标注效果显示本文的方法具有一定的普适性。

本文也对各个方法在各个数据集上的标引结果进行了分析，具体分析的方法是通过绘制不同方法在同一个数据集上的 ROC 曲线来评价的，ROC 曲线下方的面积用 AUC 值来表示，一般来说，AUC 的值越大，表明该方法在这个数据集上的分类效果越好。图 2 显示了三种分类方法在本文实验过程中的 ROC 曲线图。本文综合不同实验数据上的分类情况，得到了不同方法在同一个数据集上的 ROC 曲线图。



(a) D06 数据集　　(b) E03 数据集

图 2 三种标引方法在 IPC 分类号为的 D06、E03 数据集上的 ROC 曲线

表 4 各种分类算法在不同数据集上分类 AUC 值比较

	D06	E03
CDPLSA	0.4739	0.8578
TLPLSA	0.3107	0.9487
Ours	0.4874	0.8676

通过图 2 以及表 4 可以看出，本文方法的 AUC 值相较于 CDPLSA 和 TLPLSA，在专利数据集 D06 上本文方法的 AUC 值明显高于其他方法。在专利数据集 E03 上，本文方法的 AUC 值分别不如 TLPLSA 方法，但是 TLPLSA 在专利数据集 D06 上的分类效果却是非常差的，并且在专利数据集 E03 上，本文的方法与最优值之间的差距也不大。总体而言，本文的方法相较于其他两种衍生算法在不同数据集上的适用性和准确性上还是有比较大的优势。

通过对上述主题中词语的分析可以发现，本文提出的方法能较为准确地判别专利语料库中的共同主题、训练集中所对应的主题、测试集中所对应的特定主题。是一种非常实用的专利标引方法，适用于大量未标记专利的情形。

3 总结与展望

本文提出的方法，充分考虑到了已标引数据和未标引数据的分布情况，并建立两者之间的映射关系。它基于 PLSA 算法，演算出共同主题、训练集的特定主题、测试集的特定主题。对于这些主题词进行分析可以发现，这些演算出来的主题与训练集、测试集中所对应的 IPC 分类号所表达的含义高度的吻合，并且共同主题也能充分准确地描述训练集、测试集之间的相互关系，映射矩阵也能直观地展示出训练集

与测试集中主题之间的关系，选择相关度最高的主题作为 IPC 分类号标引的依据。在此对方法的优势与不足进行总结与展望：

(1) 该方法能适用于实际情况，特别是只有少量标引数据的情形下，本文提出的方法可以拓展专利标引的应用范围，实现大量未标记专利的标引。

(2) 该方法能使标引更加准确，通过 PLSA 模型对主题进行分析，并将专利主题分为训练集中的特定主题和测试集中的特定主题，从语义层面对专利内容进行分析，从相关度矩阵分析更加科学与便捷，使得标引更加精确。

(3) 该方法在实验中已经确定了训练集和测试集，因此共同主题、训练集和测试集的特定主题的个数比较好确定，所以标引的结果较为准确。但是如果实际情况中，不知道训练集、测试集文本的具体主题分布情况时，其标引分析的准确性将会受到影响。

(4) 该方法时间复杂度较高，因为传统 PLSA 算法采用 EM 迭代算法来进行求解，时间复杂度较高。

针对方法的优劣分析，今后还要对以下问题进行研究：

(1) 本实验选取的专利数量有限，并且训练集、测试集中只有两种大类的 IPC 号。因此在建立模型时主题个数的参数可以较为直接地确定，若是实际情况中，训练集以及测试集的相应主题个数不能轻易确定，如何找到自适应的确定主题个数的方法是下一步要研究的方向。

(2) 本实验将训练集与测试集中专利的数量设定为相近，但是实际运用中训练集的数量可能非常稀少，如何建立只通过少量的训练集就能完成快速准确地标引是下一步需要思考的问题。

(3) 传统 PLSA 算法采用 EM 迭代算法来进行求解，时间复杂度较高，使其在处理大数据上性能受到影响，如何避免 PLSA 算法时间复杂度高的缺陷，并将它广泛应用于专利标引的研究也是下一步需要解决的问题。

参考文献

[1] 李文兰 , 段晓伟 . 基于美国专利探析中国国际专利发展态势 [J]. 情报工程 , 2018, 4(4):50-61.

[2] Wartena C, Brussee R, Slakhorst W. Keyword Extraction Using Word Co-occurrence[C]. Proceedings of 2010 Workshop on Database and Expert Systems Applications (DEXA), Bilbao, Spain. New York, USA: IEEE. 2010: 54-58.

[3] 罗准辰 , 王挺 . 基于分离模型的中文关键词提取算法研究 [J]. 中文信息学报 , 2009, 23(1): 63-70.

[4] 李军锋 , 吕学强 , 周绍钧 . 带权复杂图模型的专利关键词标引研究 [J]. 现代图书情报技术 , 2015(3):26-32.

[5] 索红光 , 刘玉树 , 曹淑英 . 一种基于词汇链的关键词抽取方法 [J]. 中文信息学报 , 2006, 20(6):25-30.

[6] Noh Y, Son J W, Park S B. Keyword extraction from dialogue sentences using semantic and topical relatedness[C]. International Conference on Neural Information Processing. Springer, Berlin, Heidelberg, 2013.

[7] 丁杰 , 吕学强 , 刘克会 . 基于边界标记集的专利文献术语抽取方法 [J]. 计算机工程与科学 , 2015, 37(8):1591-1598.

[8] 刘小蝶 , 朱筠 , 晋耀红 . 中文专利中有标记并列结构的自动识别研究 [J]. 计算机工程 , 2018, 44(6):162-168+175.

[9] 章成志 . 基于集成学习的自动标引方法研究 [J]. 中国索引 , 2009, 7(2): 16-23.

[10] Chen X, Peng Z, Zeng C. A Co-training Based Method for Chinese Patent Semantic Annotation[C].

doi:10.3772/j.issn.2095-915x.2020.03.002

Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York, USA: ACM. 2012: 2379-2382.

[11] Hofmann T. Unsupervised Learning by Probabilistic Latent Semantic Analysis[J]. Machine Learning, 2001, 42(1-2):177-196.

[12] 蒋铭初 , 潘志松 , 尤峻 . 基于 PLSA 主题模型的多标记文本分类 [J]. 数据采集与处理 , 2016, 31(3):541-547.

[13] 何炎祥 , 刘续乐 , 陈强 , 等 . 社交网络用户兴趣挖掘研究 [J]. 小型微型计算机系统 , 2014, 35(11):2385-2389.

[14] 吉余岗 , 李依桐 , 石川 . 融合异质网络与主题模型的方面分预测 [J]. 计算机应用 , 2017, 37(11):3201-3206.

[15] 包翔 , 刘桂锋 , 杨国立 . 基于多示例学习框架的专利文本分类方法研究 [J]. 情报理论与实践 , 2018, 41(11):144-148.

[16] 杨奋强 , 刘玉贵 . 文本分类中基于类别概念的特征选择方法 [J]. 计算机系统应用 , 2009, 18 (10):93-96.

[17] Zhuang F, Luo P, Shen Z, et al. Collaborative dual-plsa: mining distinction and commonality across multiple domains for text classification[C]. Proceedings of the 19th ACM international conference on Information and knowledge management. ACM. 2010: 359-368.

[18] Zhuang F, Luo P, Xiong H, et al. Exploiting associations between word clusters and document classes for cross-domain text categorization[J]. Statistical Analysis & Data Mining, 2011, 4(1):100-114.