情报学报　2018 年 7 月　第 37 卷　第 7 期
Journal of the China Society for Scientific and Technical Information, July 2018, 37(7): 732-741

DOI: 10.3772/j.issn.1000-0135.2018.07.009

基于 LDA-HMM 的专利技术主题演化趋势分析——以船用柴油机技术为例

陈　伟[1]，林超然[1]，李金秋[1]，杨早立[2]

（1. 哈尔滨工程大学经济管理学院，哈尔滨　150001；2. 北京工业大学经济管理学院，北京　100124）

摘　要　如何在专利数据海洋中挖掘技术主题的研究现状、识别具有潜力的研发热点，对企业和国家来说都是至关重要的战略议题。针对目前技术主题演化趋势预测研究中存在的不足：技术创新过程中随机特征的忽视、人工分类的缺陷以及专业术语难以识别等问题，本研究提出一种组合方法，首先使用维特比（Viterbi）算法识别专利文献中的专业术语，其次利用机器学习中的隐含狄利克雷分布（LDA）算法捕捉专利文献中潜在的技术主题聚类，分析各时期技术主题的分布特征和演变规律，然后结合包含双重随机过程的隐马尔可夫模型（HMM）对未来技术趋势进行定量预测，最后以船用柴油机技术为例，应用上述组合方法分析船用柴油机技术的主题分布、演化规律及未来趋势。对比实验显示本文方法具有有效性和实用价值。

关键词　主题模型；隐含狄利克雷分布；隐马尔可夫过程；技术演化

Analysis of the Evolutionary Trend of Technical Topics in Patents Based on LDA and HMM: Taking Marine Diesel Engine Technology as an Example

Chen Wei[1], Lin Chaoran[1], Li Jinqiu[1] and Yang Zaoli[2]

(1. *School of Economics and Management, Harbin Engineering University, Harbin*　150001;
2. *School of Economics and Management, Beijing University of Technology, Beijing*　100124)

Abstract: Identifying potential research hotspots from a large number of patents is a crucial strategic issue for both enterprises and countries. In view of the problems in the current analysis of patents, such as the non-repeatability of manual classification and unrecognized specialized vocabulary in natural language processing, a combination method is proposed here as follows. First, we use the Viterbi algorithm to identify specialized terms in patent documents. Second, we introduce the LDA algorithm from machine learning to capture latent topic clusters in patent documents. Third, combining the hidden Markov model and double stochastic process, the distribution and evolution of existing technology topics are analyzed and future technical trends are predicted. Finally, this study uses marine diesel engine technology as an example of applying the above combination method to analyze the topic distribution, evolutionary pattern, and future trend of marine diesel engine technology. The experimental results prove that the proposed method shows better performance.

收稿日期：2017-12-27；修回日期：2018-07-18

基金项目：国家自然科学基金“多源异构信息交互作用下知识产权战略目标偏差的诊断与机理研究”（71704007）；国家社会科学基金项目“区域知识产权战略系统协同与产业升级研究”（14BGL007）。

作者简介：陈伟，男，1957 年生，博士生导师，主要研究方向为技术创新与知识产权管理；林超然，男，1987 年生，博士生，主要研究方向为技术创新与知识产权管理，E-mail: hcab@qq.com；李金秋，女，1989 年生，博士生，主要研究方向为技术创新与知识产权管理；杨早立，男，1986 年生，博士，主要研究方向为知识产权管理与创新评价与多属性决策。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Key words: topic model; LDA; HMM; technological evolution

1 引　言

技术主题是技术文献的主旨和核心，其演变遵循着一定的内在规律[1]，掌握技术主题的演化规律，在企业层面可缩短研发周期、节约研发费用[2]、判别技术现状并洞察行业趋势[3]，在国家层面能够明晰技术演变方向、引导产业占领未来技术高地，保护国家战略利益。因此，了解其发展历程、识别其演化路径并预测其未来热点是重要的前沿课题。在技术主题的研究中，专利文献是重要的信息来源，专利文献承载世界上 90%~95%的技术信息[4]，其内容准确翔实并具有长期性、广泛性和趋前性[5]。但专利文献数量庞大，利用科学的分析框架，对专利文献进行高效的分析是技术主题演化研究的关键所在。

为挖掘专利文献中的技术趋势，传统的方法一般利用专家经验分析代表性专利[6]，以获取其技术主题情况，这类方法无法避免专家资源稀缺、专利的代表性难以衡量、无法分析大量专利等缺陷。为避免依赖专家经验，学者尝试通过专利的分类属性作为其技术主题，例如，利用 IPC 分类号[7]、德温特手工代码[3]、专利申请人[8]等特征分析某领域内全部专利的演化特征，但专利技术主题众多，而专利属性种类有限，势必影响演化趋势分析的精确性。为更精确的划分专利技术主题，学者通过专利共现网络[9]和引用关系[10]为专利聚类，但此类方法会产生时滞，无法保证技术主题演化趋势分析的及时性。为兼顾技术主题的多样性以及趋势分析的及时性，学者使用 SAO 结构语义相似度识别[11]、主题模型[12]或主题聚类[13-14]等方式从专利等科技文献中挖掘技术主题，但中文专利中的专业词汇是识别难点，降低了这些方法的适用性。在专利技术主题演化趋势分析的手段上，学者借助技术主题的时间信息，使用灰色预测[15]、技术生命周期[2]、集对分析[16-17]、技术路线图[18]、时间序列分析[19]等方式预测技术主题演化趋势。但随机性是创新过程中的本质现象[20]，这些分析手段普遍忽视了创新过程的随机属性，在对未来技术主题演化趋势的定量预测上，忽视随机性会导致高估已有技术主题的持续性，低估新技术的爆发式增长。

为解决上述问题，本文在专利文献的处理上，使用 Viterbi 算法识别专利文献中的专业术语，弥补主题分析中分词词库缺乏专业术语的问题。在此基础上，提出一种组合方法，结合隐含狄利克雷分布（LDA）和隐马尔可夫过程（HMM）：通过 LDA 模型对海量异构专利文献数据进行主题建模，训练模型生成技术主题，避免人工标注法引起的效率和精度问题；通过引入包含双重随机过程的隐马尔可夫模型，从技术主题变化的微观角度探讨技术主题的热度变化及内容变化，预测技术趋势并做可视化展示。

2 模型构建

2.1 专利文献预处理模块

中文自然语言处理离不开分词操作，传统分词方式大多使用词库配合权重值匹配文档中的语句，通过维护庞大的词库保障分词准确性。但技术文献中存在大量专业术语，单纯使用穷举法增补词库不仅效率低，而且在宏观上也缺乏判断专业术语是否收集全面的指标。

因此，本文使用维特比算法（Viterbi Algorithm）补充语料库中未知词汇，该方法是一种动态规划算法，由 Viterbi[21]提出，配合前缀词典进行快速成词扫描，针对单句生成所有分词方式组合生成有向无环图，并利用动态规划法寻找概率最大化的切分路径，可用于挖掘与上下文无关的派生字符串。其能够在无需人工参与的条件下识别专利词汇等新词，在技术文献分词中，可有效补充标准语料库缺乏专业术语的问题。

2.2 技术主题提取模块

主题模型中 LDA（Latent Dirichlet Allocation）算法由 Blei 等[22]提出，是一种混合概率模型，通过最大化词语共现概率寻找词语聚类，利用狄利克雷分布刻画文档生成过程，并限定文档主题数量，避免 PLSI 方法过拟合以及参数过多问题。可高效提取文档隐含主题，并对文档聚类。

本文假定专利文献的技术主题服从超参数狄利克雷先验分布：

$$\mathrm{Dir}(\theta_d \mid \alpha)=\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{dk}^{\alpha_k-1} \tag{1}$$

其中，θ_{dk} 表示专利文献 d 在技术主题 k 中的分布。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

对每一个技术主题 k 生成主题词项分布 $\phi_k \sim \mathrm{Dir}(\beta)$；对每篇专利文献 d，生成主题词分布 $\theta_d \sim \mathrm{Dir}(\alpha)$，对每篇专利文献中的第 n 个词项生成主题项 $z_{dn} \sim \mathrm{Multinomial}(\theta_d)$ 和词项 $w_{dn} \sim \mathrm{Multinomial}(\phi_{z_{dn}})$。因此，本文中 LDA 似然模型可被描述为：

$$p(W \mid \alpha, \beta) = \prod_{d=1}^{D} \int p(\theta_d \mid \alpha) \prod_{n=1}^{N_d} \sum_{z_d} p(z_{dn} \mid \theta_d) p(w_{dn} \mid \phi_{z_{dn}}) d\theta_d \quad (2)$$

LDA 主题模型发挥降维作用的关键在于对异构文本潜在主题数量的准确设定，但 LDA 方法自身并不能生成最佳的主题数量。Blei[23]提出使用困惑度（Perplexity）作为确定主题数量的标准，但容易导致主题间相似度过大。Teh 等[24]提出层次狄利克雷过程（Hierarchical Dirichlet Processes），其使用非参数模型自动训练得到主题数，但方法运算效率较低，针对大规模文本分析时很难保证迭代精度。考虑到模型的泛化能力及主题抽取效果，本文使用考虑了困惑度和主题间相似性的 Perplexity-Var 方法[25]计算最优主题数，该方法通过主题的散度（D_{JS}）衡量主题的结构稳定性，并惩罚过多的过量主题，在确保主题间区别最大化的前提下，尽量减少主题数。最后，采用 Heinrich[26]的参数估计方法，设定 $\alpha = 50/K$，$\beta = 0.1$。使用吉布斯抽样（Gibbs Sampling），得到主题集合 $K = \{k_1, \cdots, k_h\}$，以及每项专利的技术主题归属 $D_k = \{j_1, \cdots, j_n\}$。

2.3 技术主题演化趋势分析模块

技术主题的演化存在两种内在动力：一是在技术更迭的过程中，历史研究成果启发和支撑新技术思想的产生，由于缺少记录载体，这一过程为不可观测的隐藏序列；二是在第一种动力的推动下，随着研发环境改变以及非期望研发产出的显现，研发人员不断调整研发预期，进而改变研究成果，专利文献将研究成果有效记录，成为可观测序列。后者构成前者的微观基础，前者是后者的宏观表现。因此，技术主题的演变可被视为这两个过程的叠加。

本研究使用隐马尔可夫模型刻画上述技术主题的演化过程。通过推断隐马尔可夫模型中的状态转移矩阵及初始状态的概率分布，确定技术主题演变中主题间的混淆矩阵及转移矩阵，进而确定技术主题的演变历史和未来演化趋势。基于以上思想，本文构建隐马尔可夫模型如下：

（1）隐藏状态随机转移序列集合：$S = \{s_1, \cdots, s_h\}$，其中 h 为 LDA 模型中的主题数，设随机过程生成的隐状态序列为 $Q = \{q_1, \cdots, q_t\}$，其中 $q_t \in S$。

（2）状态转移概率分布为 $A = \{a_{ij}\}$，其中 $a_{ij} = P\{q_{t+1} = S_j \mid q_t = S_i\}$，$1 \leqslant i, j \leqslant N$，且满足 $a_{ij} \geqslant 0$，$\sum_{j=1}^{N} a_{ij} = 1$，其表示研发过程中，研发人员研发主题从状态 S_i 转移到 S_j 的概率。

（3）状态为 S_i 时，观测变量的概率分布为：$B = \{b_i(v)\} = \{f\{Q_t = v \mid q_t = S_i\}\}$，其中 Q_t 为第 t 种观测随机变量，观测到序列为 $O = \{o_1, \cdots, o_t\}$，本文中观测序列为历年各技术主题比重。

（4）系统初始状态的概率分布为 $\pi = \{\pi_i, 1 \leqslant i \leqslant N\}$，其中 π_i 为出现状态 S_i 的概率。一般来说，技术主题间主题词共现次数越高，主题间的关系越密切，也越容易产生技术主题的迁移和演进，因此，本文中使用主题词共现数归一化矩阵作为 π_i 的初始迭代值。

上述模型可表示为图 1。

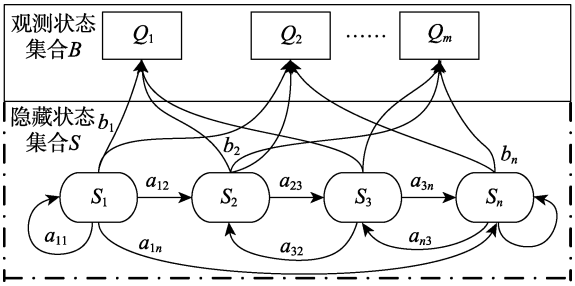


图 1 隐藏状态转移序列与观测序列关系

设定模型训练初始值：$O = Q = \{pt_1, \cdots, pt_t\}$，使用 Baum-Welch 算法[27]对上述 HMM 模型进行参数估计，得到单一最优状态序列，并根据 $\hat{O}_{t+k} = \sum_{j=1}^{N} A^k(i, j) E(b_j(v))$ 获得 k 年后的研发主题结构。

综上所述，本文研究模型设定如图 2 所示。

3 案例分析

为验证本文提出方法的可行性和有效性，基于本文第 2 节所构建的模型，本节对船用柴油机的技术主题演化情况做出分析。目前，中国已经逐渐成为具有广泛海洋利益的大国，向海洋进军是中国地

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

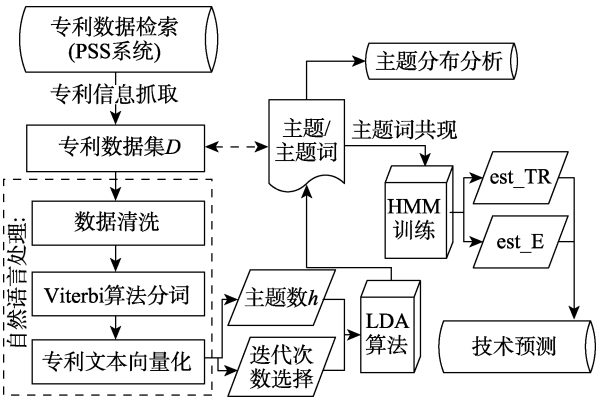


图 2　本文研究思路

缘战略的重点[28]，而船舶是获取战略资源的重要载体。近年，我国船舶工业技术水平和市场份额不断提升，但船用柴油机作为船舶动力的来源，我国却缺乏对其核心技术的战略布局。面对技术保护为导向的国际环境，国家对关键技术领域进行政策倾斜，并科学的引导研发资源是十分必要的。因此，本节以船用柴油机技术为例，探讨船用柴油机技术主题现状及其演化趋势。

3.1　数据来源及预处理

3.1.1　专利检索

本文使用“国家知识产权局专利检索及分析网”（www.pss-system.gov.cn）中的专利文献数据库检索船用柴油机技术相关专利，检索时间为 2017 年 10 月 27 日，得到与船用柴油机相关专利 6083 项，申请时间跨度为 1985 年 4 月至 2017 年 8 月。

使用 Python 的 Urllib 库[29]批量抓取专利相关字段（名称、摘要、申请日期、分类号、申请号等），实现自动化批量数据采集整理。检索得到的专利主要集中于：F02（燃烧发动机）、F01（一般机器或发动机）、B63（船舶或其他水上船只）、C10（石油、煤气及炼焦工业）等。

3.1.2　分词

本案例中，Viterbi 算法发现新词 2300 余个，例如：推杆、推流器、浮筏、脱硝、脱硫、喷油器、电磁阀、冷凝器、凸轮等，对专利文献分词结果抽检可发现分词效果较为理想。处理后，定义 $D=\{d_1,\cdots d_n\}$ 为专利数据集，其中 d_n 为第 n 项经分词处理后的专利文献文本向量。

3.2　专利文献的主题提取

LDA 模型需设定其运行参数。首先，LDA 是一种机器学习算法，其学习效果与迭代次数密切相关[30]。本案例中，随着迭代次数增加，模型迅速收敛，迭代至 500 次之后，收敛效果并无明显区别（图 3），为加快运算速度，本文将迭代次数设定为 500 次。其次，本文计算了 5 至 150 个主题数时的 Perplexity-Var 值，当划分为 42 个主题时，Perplexity-Var 值最小，如图 4 所示。因此设定主题数为 42。

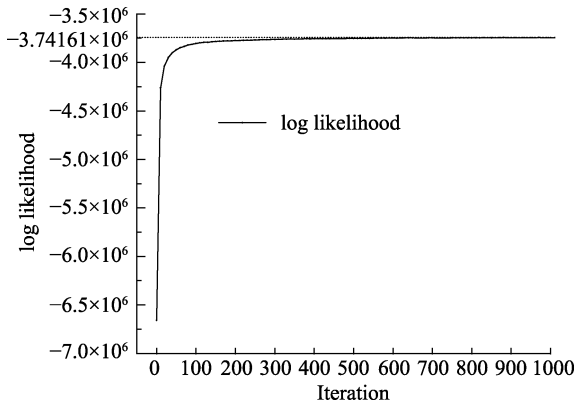


图 3　迭代次数对机器学习效果的影响

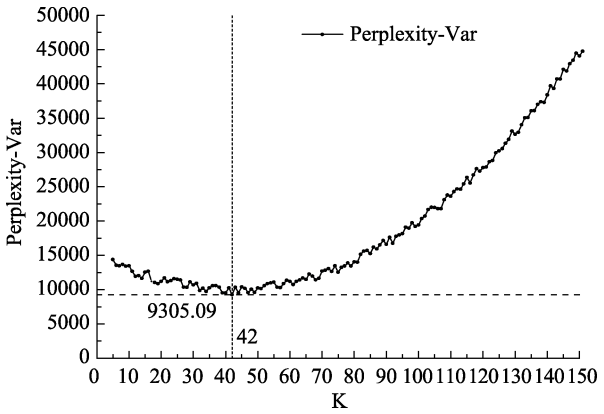


图 4　不同主题数的 Perplexity-Var 值

使用 Python 中 LDA 库计算得到主题信息，导出每个主题的主题词。根据主题词情况过滤掉与船用柴油机技术无关的主题聚类（4 个），去掉由虚词组成的聚类（3 个），剩余 35 个主题。这些主题之间技术领域边界清晰，分类效果较为理想。为便于指代，本文将各主题编号并命名，如表 1 所示。

为评估 LDA 模型提取技术主题的正确性，本文按照各主题中专利数量的 10%抽取专利样本，各主题的正确率如图 5 所示，正确率均值为 89.90%，总体效果较好。

基于上述结果，绘制技术主题的数量和比重趋势热度图（图 6）。其中，观察图 6b 中技术主题的首次出现时间以及年度变化趋势可知，主题种类随着

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 1 主题关键词及命名

No.	KW1	KW2	KW3	KW4	KW5	ID
1	壳体	固定	内部	挡板	设置	壳体固定
2	固定	柴油机	安装	定位	工具	定位安装
3	驱动	液压	装置	马达	油缸	驱动液压装置
4	密封	柴油机	阀体	柱塞	针阀	柴油机密封
5	装置	发电	蓄电池	太阳能	风力	发电/蓄电装置
6	涡轮	增压器	柴油机	进气	增压	涡轮增压
7	气缸套	柴油机	工艺	合金	焊接	气缸套工艺
8	控制	调节	系统	组件	船舶	控制调节系统
9	输出	输入	齿轮	齿轮箱	离合器	齿轮箱离合器
10	发动机	柴油	燃油	节能	动力	节能柴油机
11	系统	推进	发电机	船舶	发电机组	电推进系统
12	制备	柴油	复合	添加剂	燃料	复合燃料
13	船体	推进	螺旋桨	推进器	航行	螺旋桨推进器
14	燃料	气体	发动机	喷射	燃烧室	发动机燃烧室
15	系统	催化剂	SCR	柴油机	废气	废气催化
16	电源	电路	输出	开关	连接	电路控制
17	处理	装置	船舶	废气	脱硫	废气处置
18	轴承	转子	叶片	旋转	叶轮	叶片与轴承
19	添加剂	润滑	烷基	润滑油	润滑剂	润滑添加剂
20	柴油机	安装	结构	本体	船用	安装
21	连接	支撑	固定	安装	设置	支撑与固定
22	润滑油	柴油机	机油	润滑	注油	机油
23	测量	柴油机	传感器	计算	步骤	传感器
24	船舶	主机	试验	仿真	模拟	仿真模拟
25	余热	回收	系统	装置	利用	余热回收
26	冷却	柴油机	水泵	冷却水	海水	海水冷却
27	齿轮	柴油机	动力	传动	机构	动力与传动
28	曲轴	连杆	柴油机	船用	曲柄	曲轴曲柄
29	加工	运动	机床	零件	夹紧	零件加工
30	平台	移动	装置	作业	升降	作业平台
31	船体	装置	收集	垃圾	清淤	清淤
32	燃油	柴油机	加热	管路	冷却	管路加热
33	监控	柴油机	信号	传感器	控制	监控
34	活塞	气缸	发动机	内燃机	柴油机	活塞技术
35	高压	压力	柴油机	系统	共轨	高压共轨

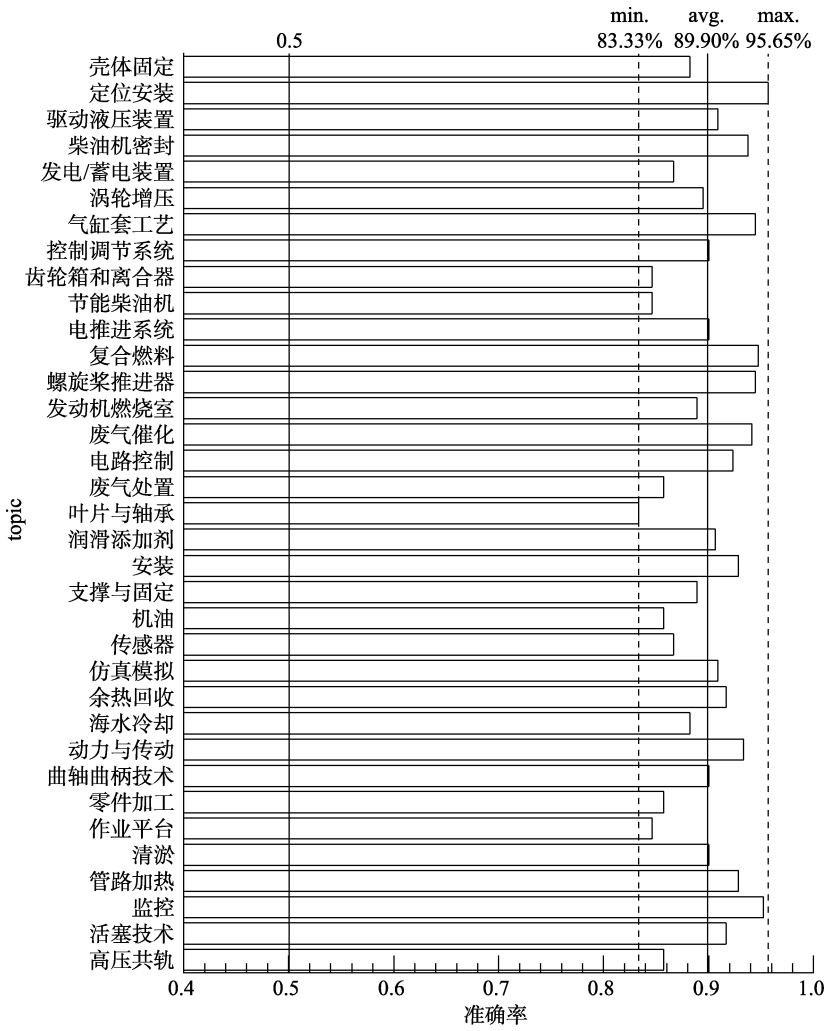
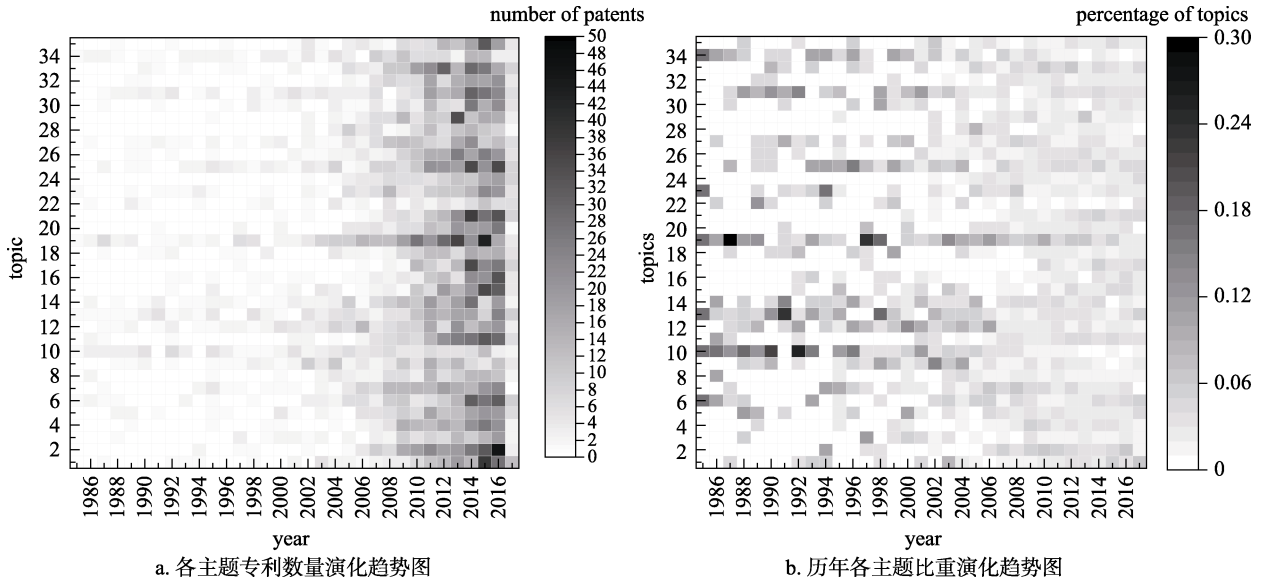


图 5 各主题分类的准确率

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net



a. 各主题专利数量演化趋势图　　b. 历年各主题比重演化趋势图

图 6　各主题专利数量与比重演化趋势图

研究逐渐丰富不断得到细分，早期仅存在主题 6（涡轮增压）、10（节能柴油机）、13（螺旋桨推进器）、19（润滑添加剂）、23（传感器）及 34（活塞技术）。随着柴油机的电控技术逐渐兴起，产生主题 8（控制调节系统）、11（电推进）和 16（电路控制）等。船用柴油机技术发展至 2000 年左右，技术主题演化出现三种趋势：一是柴油机能效出现瓶颈，迫使柴油机加工精度提高，推动相关领域研发，产生主题 21（支撑与固定）及 28（曲轴曲柄技术）；二是随着电脑仿真技术的逐渐成熟，针对船用柴油机的仿真模拟技术（主题 24）的探索也逐渐增加；三是随着科技进步以及环保压力逐渐增大，近年针对废气处置（主题 17）和高压共轨（主题 35）的研究增多。

3.3　主题演化路径分析

主题间之间主题词越相似，主题间出现混淆或转移概率越高。因此，本文采用主题词共现分析方法，统计 35 个技术主题中前 40 个主题词的共现数量作为主题间相似程度的表征，构建 35×35 共现词数对称矩阵，并绘制热度图（图 7a）。由于各主题与自身共现程度最高，因此可观察到明显的对角线，其余区域由于共现词数不同产生深浅不一的涨落。将归一化后的共现词数矩阵（图 7b）作为 HMM 模型初始混淆矩阵及转移矩阵，导入本文构建的技术主题演化隐马尔可夫模型中，利用 Baum-Welch 算法的训练 HMM 模型，获得最优混淆矩阵（图 7c）和转移矩阵（图 7d）。

混淆矩阵代表某个隐藏的状态被观察为某种可观测状态的概率，在本案例中，这一概率衡量柴油机技术主题之间在研发过程中发生迁移的门槛阻隔，并体现为单一技术主题在研发过程中发生改变的方向和可能性，其热度图（图 7c）中深色方块代表更容易在研发过程中迁移的技术主题，图中可见总体上大多数主题之间混淆少、研发内容独立程度高、迁移门槛高，即大多数技术主题之间存在壁垒，不易发生转移。同时，少数主题之间存在不同程度的混淆可能性，本文将其中混淆概率超过 10%的主题绘制成混淆关系网络图（图 8），其中连接线粗细代表混淆可能性。

可见，研发主题较容易转移至涉及多部件的技术主题（如涡轮增压、驱动液压装置和壳体固定），较容易从门类狭窄的主题中迁移出（如发动机气缸套、零件加工和活塞技术）。此外，一些技术主题是技术演化中重要的桥梁和中介，是技术主题发展和变迁的关键环节，例如，安装技术是活塞技术和清淤技术演变至壳体固定技术的关键节点，柴油机密封研究也为零件加工和壳体固定研究的演变提供了技术层面的过渡。起到类似作用的主题还有电推进系统和支撑固定技术。混淆特征揭示了船用柴油机技术主题的潜在研发演化路径，便于微观研发部门在进行科研目标规划时能够科学选择技术路线。

转移矩阵代表在研发的宏观历史进程中主题间可能发生转移的概率，总体上看，图 7d 中对角线仍然较为明显，表明大多数主题有保持自身研究趋势的稳定性。从单一主题视角观察，可发现主题间研发趋势转移特征存在差异。本文将转移概率大于

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

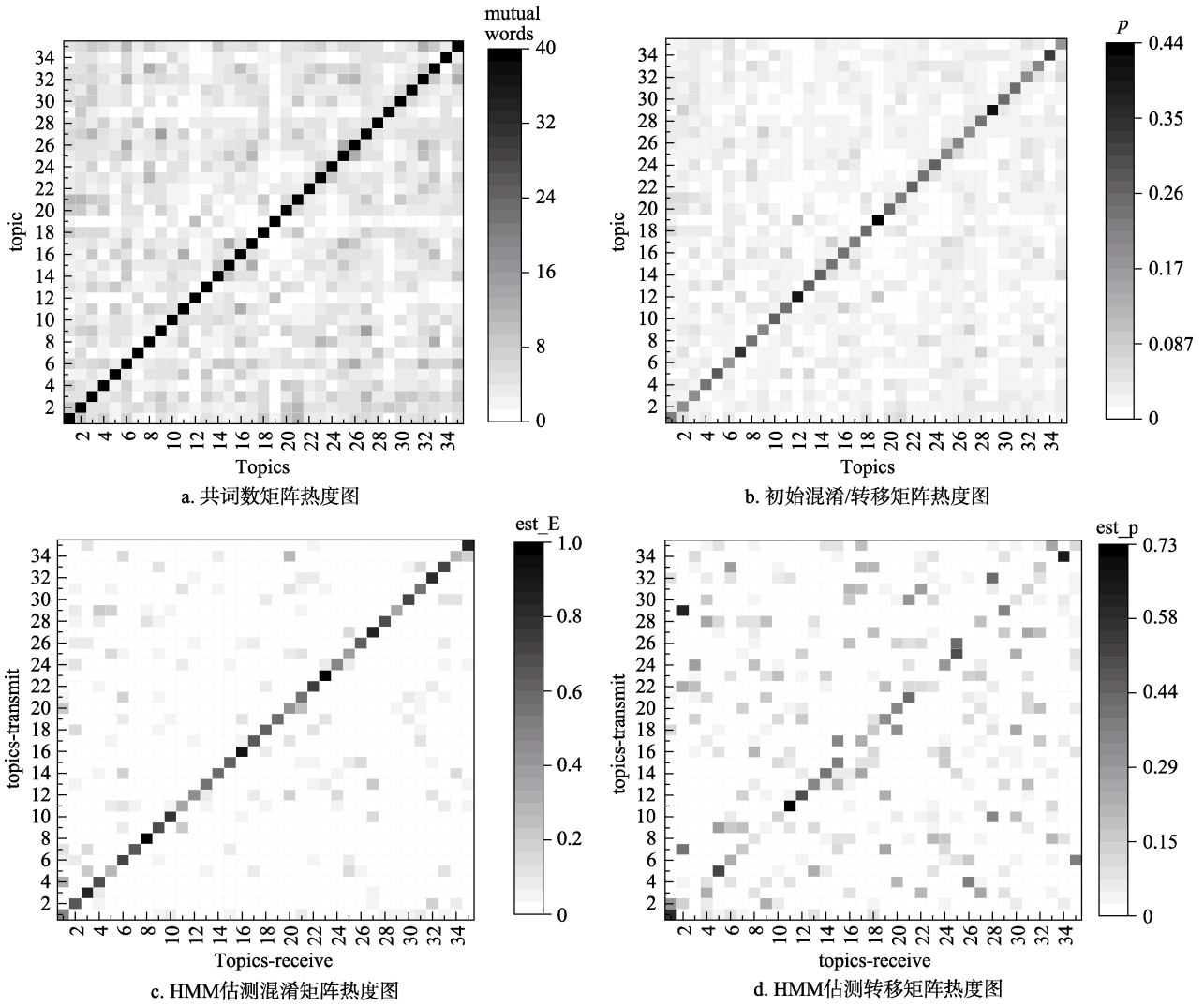


图 7 HMM 模型参数训练过程矩阵热度图

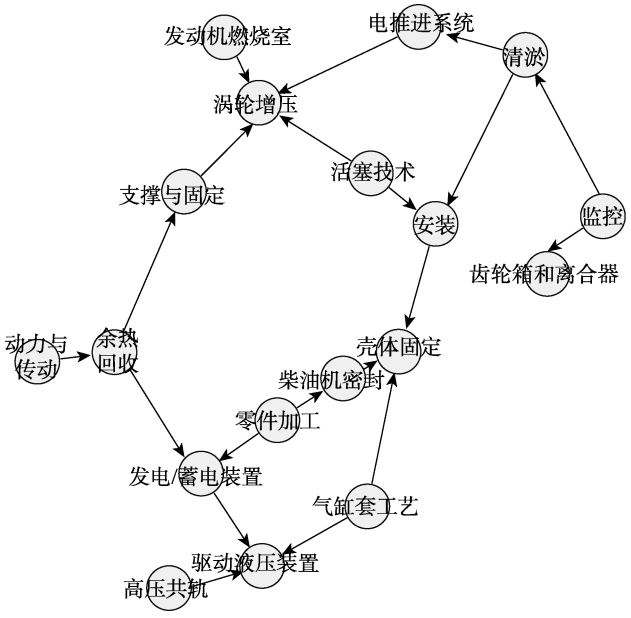


图 8 主题间的混淆关系

15%的技术主题转移绘制关系特征图（图 9）。

其中，技术主题演化中保持自身研发方向能力较强的主题包括：电推进系统（0.7298）、活塞技术（0.6473）、壳体固定（0.5876）等，这些技术主题在演化过程中一直保持着较高的热度，是柴油机相关研发过程中的热门技术节点。此外，技术主题演化过程中转移流失比重最大的主题是“监控”,其主要流向涡轮增压、废气处理等方面，这是由于柴油机监控功能的网络化、数字化趋势，单纯使用仪器仪表的模拟信号监控的研发逐渐减少。热度增加最多的技术主题是“定位安装”，其研发力量主要来自于“零件加工”和“气缸套工艺”,这是由于柴油机元件精度要求不断提升，零件加工、机油润滑等领域内的研发逐渐遇到技术瓶颈，对元器件精度的要求逐渐转向柴油机的定位、测量和校准等方向。同时技术进步也使原本需要处理的废物转变为可利用资源，

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

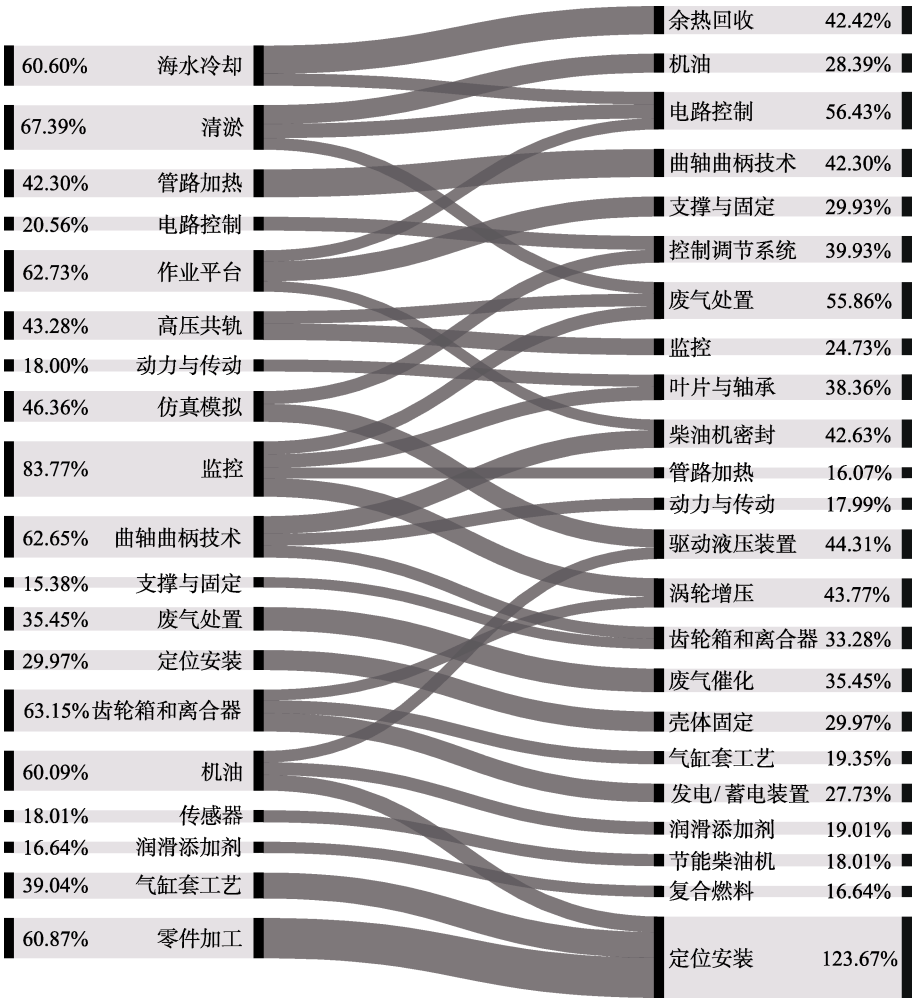


图 9　隐马尔可夫估测主题间转移概率

如海水冷却技术逐渐被余热回收技术所替代。这些技术转移过程展现了船用柴油机相关技术研发的历史变迁，并能为预测未来技术发展提供判断依据。

由于本研究进行时，2017 年专利文献尚未完全公布，因此，在对未来船用柴油机技术主题演化趋势的预测上，本文使用 2016 年专利数据作为预测基期，将估测混淆矩阵和转移矩阵参数导入 Matlab 中 HMM 模块，得到 2017—2020 年，船用柴油机各技术主题演变的隐马尔可夫预测结果，如图 10 所示。

隐马尔可夫预测结果中，活塞技术与曲轴曲柄技术的比重提升较大，这是由于曲柄连杆机构控制活塞的同步开闭日渐成为研发热点；节能减排仍然是船用柴油机的主要研发方向，其中研发重点从涡轮增压（降低排放、回收能量）转向高压共轨技术，该技术能够满足柴油机在不同工况条件下对喷油压力的需求，可柔性调节喷油率，改善燃烧过程，提高柴油机的工作性能并降低有害排放。可见，在资源约束及环保压力下，节能技术将成为重要的研发方向。

定位安装与支撑固定技术的研发比重下降明显，这一技术对柴油机性能的改善受制于柴油机的工艺水平，因此定位安装技术研发热潮很快进入低费效比阶段。两者的研发会逐渐转向壳体固定技术，该技术可有效降低船体结构性摩擦阻力并具有更好的降噪效果。

3.4　效果评价

为评估本文 LDA-HMM 方法的有效性，本文利用第 3.1 节所获得的专利文献数据，运用 IPC 分类及灰色预测（Grey Forecast）作为替代方法获得演化趋势，并对比效果。其中，使用 IPC 分类号作为 LDA 的替代方案；使用灰色预测作为 HMM 的替代方案。将替代方案与原方案两两组合，共对比四种技术主题提取及演化趋势分析方案：LDA-HMM、LDA-GreyForecast、IPC-HMM 以及 IPC-GreyForecast。通过对比 2017 年（截至 10 月 27 日）各技术主题比重

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

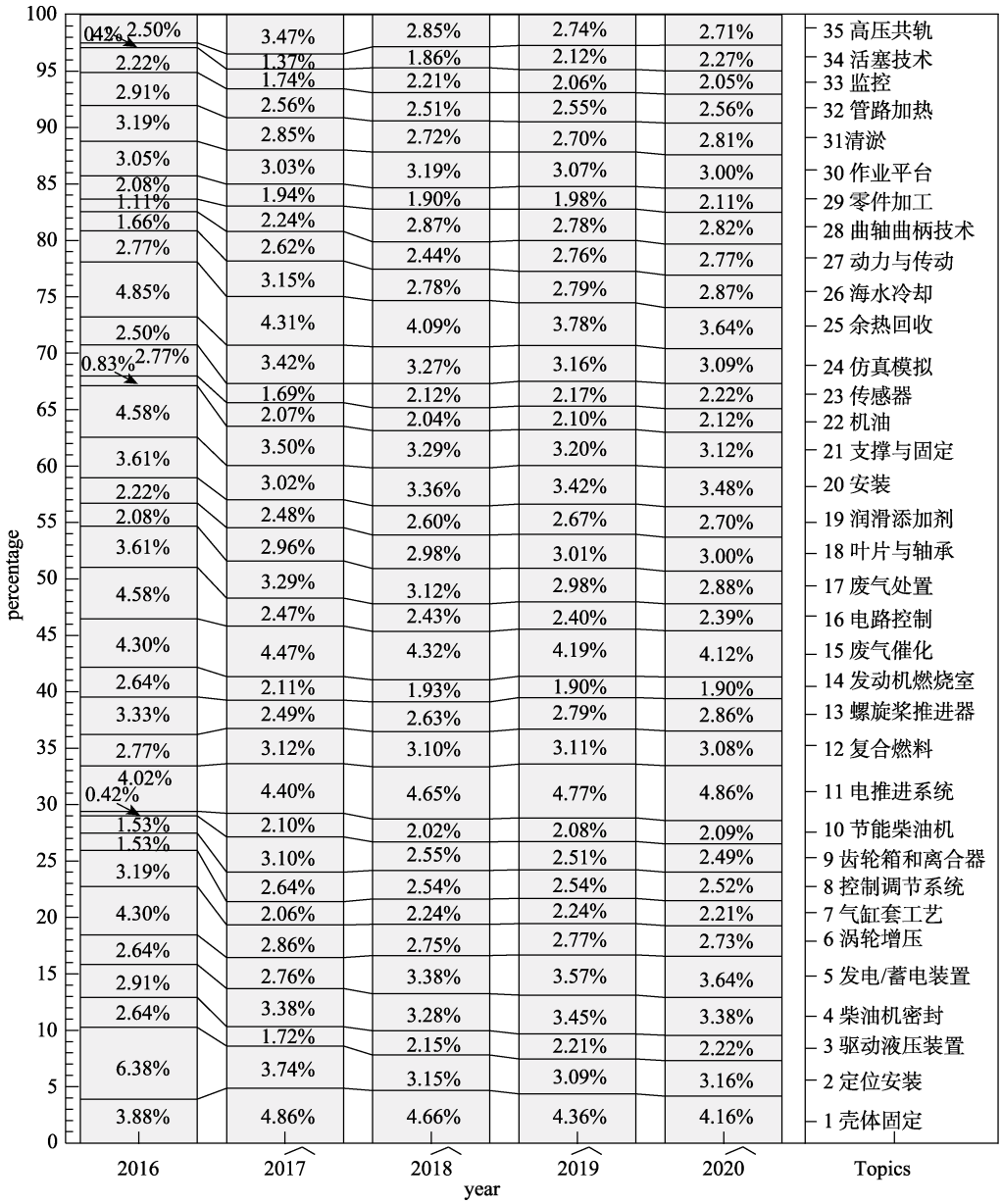


图 10 2017—2020 年技术主题演化趋势预测

预测值与实际值的平均误差，衡量四种方案的优劣，结果如表 2 所示。

表 2 各方案平均误差统计表

方案名	平均误差
LDA-HMM	0.01206
LDA-GreyForecast	0.02180
IPC-HMM	0.01945
IPC-GreyForecast	0.02010

可见，LDA-HMM 误差最小，预测效果最好。并且，IPC 分类号能够提供的技术主题信息有限，仅能将专利划分为 11 个主题分类，在主题分类精度上劣于 LDA-HMM 模型。

4 结论与展望

本文结合 LDA 主题模型以及隐马尔可夫模型，尝试提出一种新的技术主题挖掘与预测技术，实现无需专家经验的专利信息高效无监督聚类，并快速挖掘其中隐含的技术主题信息，获得关键的技术节点及技术演化趋势，为专利计量提供新途径。

在我国船用柴油机技术发展趋势案例中，该方法将船用柴油机技术领域划分为 35 个技术主题，其中，围绕涡轮增压、驱动液压装置、壳体固定的技术得到深入的开发，针对节能减排、曲柄及活塞新

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

特性的研发是未来技术创新和产品竞争的重点，而我国研发力量在这些领域积累不足，创新能力上处于弱势地位，如何抓住机会形成优势，是我国船用柴油机技术面对的挑战，应引起企业技术研发机构及政府有关部门的重视。

案例分析发现，该方法聚类效果较好，相较同类技术预测误差更小，充分说明本文方法是有效且实用的。未来的研究包括：进一步提高技术主题识别的准确性；降低模型算法复杂性，加快迭代效率，以实现针对更大规模专利数据集的分析工作。

参 考 文 献

[1] Kelly K. What technology wants[M]. Penguin Books, 2010: 71-74.

[2] 谢志明, 张媛, 贺正楚, 等. 新能源汽车产业专利趋势分析[J]. 中国软科学, 2015(9): 127-141.

[3] 韩震, 沈君, 曲莎莎. RFID 技术趋势及竞争态势的专利计量分析[J]. 科研管理, 2013(7): 11-16.

[4] 中国科学院综合计划局, 中国科学院国家科学图书馆成都文献情报中心. 中国科学院专利分析报告[R]. 成都: 中国科学院, 2015.

[5] 林岩. 基于专利数据的知识计量研究评述[J]. 科技管理研究, 2008(9): 91-93.

[6] 余江, 陈凯华. 中国战略性新兴产业的技术创新现状与挑战——基于专利文献计量的角度[J]. 科学学研究, 2012(5): 682-695.

[7] 刘云, 刘璐, 闫哲, 等. 基于专利计量的全球碳纳米管领域技术创新特征分析[J]. 科研管理, 2016(S1): 337-345.

[8] 刘云, 夏民, 武晓明. 中国最大 500 家外商投资企业在华专利及影响的计量研究[J]. 预测, 2003(6): 19-23.

[9] 丁堃, 曲昭, 张春博. 比较视角下的中美银行专利计量分析和创新对策研究[J]. 科研管理, 2014(9): 138-146.

[10] Magri A, Giovannini F, Connan R, et al. Nutrient management from biogas digester effluents: a bibliometric-based analysis of publications and patents[J]. International Journal of Environmental Science and Technology, 2017,14(8): 1739-1756.

[11] 李欣, 王静静, 杨梓, 等. 基于 SAO 结构语义分析的新兴技术识别研究[J]. 情报杂志, 2016(3): 80-84.

[12] Figuerola C, Marco F, Pinto M. Mapping the evolution of library and information science (1978-2014) using topic modeling on lisa[J]. Scientometrics, 2017,112(3): 1507-1535.

[13] Hu B B, Dong X L, Zhang C W, et al. A lead-lag analysis of the topic evolution patterns for preprints and publications[J]. Journal of the Association for Information Science and Technology, 2015, 66(12): 2643-2656.

[14] Jiang H C, Qiang M S, Lin P. Finding academic concerns of the three gorges project based on a topic modeling approach[J]. Ecological Indicators, 2016, 60: 693-701.

[15] Li W W. Application of grey prediction theory to forecast technology input within the Chinese high-tech industries[C]// Proceedings of the 3rd International Conference on Advanced Computer Control. IEEE, 2011: 88-92.

[16] 李柏洲, 李新. 基于集对分析的企业技术依赖预警及其演化趋势测度[J]. 运筹与管理, 2015, 24(2): 262-271.

[17] 黄鲁成, 成雨, 吴菲菲, 等. 关于颠覆性技术识别框架的探索[J]. 科学学研究, 2015, 33(5): 654-664.

[18] 李欣, 黄鲁成. 技术路线图方法探索与实践应用研究——基于文献计量和专利分析视角[J]. 科技进步与对策, 2016, 33(5): 62-72.

[19] Heo G E, Kang K Y, Song M, et al. Analyzing the field of bioinformatics with the multi-faceted topic modeling technique[J]. BMC Bioinformatics, 2017, 18(Suppl 7): 251.

[20] 官建成. 产品创新扩散中的随机现象[J]. 中国管理科学, 1994(3): 44-50.

[21] Viterbi A. Viterbi algorithm[M]. John Wiley & Sons, 2003: 6246.

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

[23] Blei D. Probabilistic topic models[J]. Communications of the ACM, 2012, 55(4): 77-84.

[24] Teh Y W, Jordan M I, Beal M J, et al. Sharing clusters among related groups: hierarchical Dirichlet processes[C]// Proceedings of the Neural Information Processing Systems Conference, 2005: 1385-1392.

[25] 关鹏, 王曰芬. 科技情报分析中 LDA 主题模型最优主题数确定方法研究[J]. 现代图书情报技术, 2016(9): 42-50.

[26] Heinrich G. Parameter estimation for text analysis[R/OL]. http://rakaposhi.eas.asu.edu/f12-cse571-mailarchive/pdflwcW7WccCL.pdf.

[27] Welch L. Hidden Markov models and the Baum-welch algorithm[J]. IEEE Information Theory Society Newsletter, 2003, 53(4): 10-13.

[28] 刘新华. 中国海洋战略的层次性探析[J]. 中国软科学, 2017(6): 1-13.

[29] Hetland M. Network programming[M]// Beginning Python. Springer, 2017: 273-287.

[30] AlSumait L, Barbará D, Domeniconi C. On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking[C]// Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2008: 3-12.

（责任编辑　魏瑞斌）

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net