doi:10.3969/j.issn.1000-7695.2023.8.003

基于专利文本挖掘的产业关键共性技术识别与应用研究

胡 凯[1]，谢 芬[1]，杨滨瑜[2]，胡新杰[3]，刘汉霞[4]

（1. 中南民族大学经济学院，湖北武汉 430074；
2. 南开大学经济与社会发展研究院，天津 300071；
3. 中国科学技术大学管理学院，安徽合肥 230022；
4. 武汉纺织大学外经贸学院，湖北武汉 430202）

摘要：突破产业关键共性技术掣肘，首先需要科学识别兼具关键性和共性的产业技术，但目前国内关于关键共性技术识别的研究非常有限。遵循专利文献挖掘思路，兼顾主客观标准，运用大样本量数据集，研究提出一种产业关键共性技术识别方法，为政府制定有针对性的创新支持政策提供决策参考。以我国高校（2001—2020 年）和上市公司（2006—2020 年）的海量专利作为研究样本，通过构建 LDA 主题模型、计算专利关键性得分和共性得分，以及运用相似度分析将选择结果与工信部《产业关键共性技术发展指南（2017 年）》进行相似度匹配，来识别出我国产业关键共性技术。结果显示：无论是高校还是上市公司，属于关键共性技术专利的比例均不高，二者申请专利中的比例分别为 14.7%、10.2%，授权专利中的比例分别为 13.8%、10.1%，高校专利占比略高于上市公司；而得分位居前列的技术主题如“电路芯片耦合”“能效模型图像”“系统数据模块”等与集成电路、人工智能、大数据等当前广受关注的关键共性技术非常一致，表明所用方法是有效、可取的。

关键词：产业关键共性技术；关键性；共性；专利；文本挖掘；技术主题；阈值选择

中图分类号：N99；G250.252；G301 **文献标志码：**A **文章编号：**1000-7695（2023）8-0021-11

Research on Identification and Application of Industrial Key General Technology Based on Patent Text Mining

Hu Kai[1], Xie Fen[1], Yang Binyu[2], Hu Xinjie[3], Liu Hanxia[4]

(1.School of Economics, South-Central Minzu University, Wuhan 430074, China;
2. College of Economics and Social Development, Nankai University, Tianjin 300071, China;
3.School of Management, University of Science and Technology of China, Hefei 230022, China;
4.College of International Business and Economics, Wuhan Textile University, Wuhan 430202, China)

Abstract: To break through the constraints of industrial key generic technology, it is necessary to scientifically identify industrial technologies that are both critical and general. But there is few research on the identification of key generic technology in China. Therefore, according to the idea of patent literature mining, and taking into account the subjective and objective criteria, this paper uses a large sample size dataset to propose a method for identifying industrial key generic technology, in order to provide a reference for the government to formulate targeted innovation support policies. It takes the massive patents of China's universities (2001-2020) and listed companies (2006-2020) as the research samples to identify the industrial key generic technology of China, by constructing the LDA theme model, calculating the critical and generic patent scores, and using the similarity analysis to match the results with the Industrial Key General Technology Development Guide (2017) issued by the Ministry of Industry and Information Technology. The results shows that the proportion of key general technology patent is both not high in universities and listed companies, of which the proportion of patent applications is respectively 14.7% and 10.2%, and the proportion of authorized patents is 13.8% and 10.1%; the proportion of university patents is slightly higher than that of listed companies. The top scoring technical themes such as "circuit chip coupling", "energy efficiency model images" and "system data modules" are highly consistent with key generic technology of current interest such as integrated circuits, artificial intelligence and big data, indicating that this method is effective and desirable.

Key words: industrial key general technology; key; commonality; patents; text mining; technical topics; threshold selection

收稿日期：2022-08-19，**修回日期：**2022-11-10

基金项目：国家社会科学基金一般项目“产业关键共性技术研发的财政激励机制优化研究”（19BJL079）

基础材料、关键工艺、核心部件、系统集成等方面的关键共性技术供给不足已经成为制约我国产业竞争力提升的重大问题。新型冠状病毒感染疫情叠加俄乌冲突引发的全球供应链断裂风险、美国对中国实施“脱钩断链”威胁等外生冲击，迫切要求中国突破关键共性技术掣肘，实现科技自立自强。我国“十四五”规划纲要提出，为提升技术创新能力，要集中力量整合提升一批关键共性技术平台，完善技术创新体系。关键共性技术是处于竞争前阶段并具有通用性的技术[1]。关键共性技术面临市场失灵风险，因而需要政府干预[2]。政府推动关键共性技术研发，首先需要明确哪些技术为关键共性技术，为此需要采用科学合理的方法来加以识别[3]；其次需要准确了解产业关键共性技术发展现状，为此需要运用科学方法、针对大样本数据来分析[4]。通过识别哪些技术为产业关键共性技术、分析典型部门的产业关键共性技术供给现状[5]，能够为政府优化创新资源配置提供着力点、为政府破解“卡脖子”难题提供决策基础[6]。国内学者对共性技术识别的研究较多，但对关键共性技术识别的研究非常有限，仅有少数学者采用专利文献挖掘方法识别了中国机器人和数控机床领域的关键共性技术、国际新材料领域的关键共性技术。因此，本研究以我国高校和上市公司的海量专利作为研究样本，遵循专利文献挖掘思路，探索采用试错方法选择阈值，对我国产业关键共性技术进行识别。

1 文献回顾

1.1 关键共性技术的界定

关键共性技术是关键技术与共性技术的交集。“共性技术”（generic technology）这一概念面世较早且广受关注。自1981年由Granberg等[7]提出后，美国经济学家Tassey[8]在学术上对共性技术进行了界定：共性技术介于基础研究和应用研究之间，起着承上启下的作用，是一种能广泛应用的技术，通过共性技术的开发能推动众多专项技术的商业化应用。共性技术具有产业关联性强、经济社会效益显著等特征[9]。根据对国民经济的影响程度，共性技术可以分为基础性共性技术、一般共性技术和关键共性技术[10]。其中，关键共性技术建立在共性技术的基础之上，是关键的共性技术，介于基础技术和专有技术之间[11]。“关键共性技术”是具有中国特色的概念，国外与之接近的概念是“通用目的技术”（general purpose technology），指能够影响整个经济、具有划时代意义的技术，比如蒸汽机、铁路、电力、电脑、因特网等。相对而言，关键共性技术更为多样化、具体化，比如《产业关键共性技术发展指南（2017）》共提出需要优先发展的产业关键共性技术174项，包括全数字高档数控系统技术、智能网联汽车技术、集成电路专用设备及材料技术等。

1.2 关键共性技术识别方法

随着数理统计、信息情报学等理论工具的引入，理论界对关键共性技术或共性技术的识别研究逐步深入，先后形成了4种主要的识别方法。

1.2.1 基于专家经验的识别方法

基于专家经验的识别方法是借助领域专家经年积累的知识和实践总结的经验，基于当前技术、经济、环境和社会等发展情况，对未来可能产生巨大经济和社会效益的关键共性技术进行识别，主要包括德尔菲法、层次分析法、主成分分析法、专利地图分析法和技术路线图法等，其中德尔菲法使用最为广泛。使用基于专家经验的识别方法，英国率先实施了国家技术预见计划、成立技术预见专家指导小组，开发了一套较为成熟的产业共性技术预见体系[12]。袁思达[13]首次在我国使用德尔菲法对能源技术进行研究，为国家关键共性技术的选择和优先发展技术领域的确定提供了参考。曹旭华等[14]使用德尔菲法，从需求牵引的角度详细研究了浙江省新材料行业的关键共性技术。张乔木[15]通过构建技术预见指标体系，对备选技术进行了3轮德尔菲调查，最终形成山西省新材料行业13项重点关键共性技术。

基于专家经验的识别方法充分发挥专家的知识储备与趋势把握优势，具有操作灵活、针对性强等优点，能够将行业关键共性技术研究水平、预期实现时间、重要程度等关键信息揭示出来，便于决策者了解技术现状，为决策提供充分信息。但该方法的缺陷也很突出：一是其有效性依赖于专家的知识水平和判断力，但是，专家稀缺、专家权威性无法衡量和结果无法验证等问题使得专家识别效果难以客观评价；二是专家评价涉及主观判断，专家们难以对不同选项的权重赋值达成完全共识。

1.2.2 基于指标评估的识别方法

基于指标评估的识别方法主要是通过界定关键共性技术的关键特征，建立系统指标体系来开展识别工作[16]。黄鲁成等[17]指出产业共性技术具有核心性、广泛性和效益性，并以此为基础构建产业共性技术识别概念框架。Lee[18]等提出了一种机器学习方法，用于构建多个即时的专利指标来识别早期的新兴技术。王燕鹏等[19]利用文献知识聚类理论、复杂网络理论和结构洞理论等，构建了新兴技术识别的分析框架。Liu等[20]指出新兴技术具有持续性、广泛性和效益性特性，并通过这3个特性构建了一

个三维评估框架。江娴等[21]通过梳理多个学者对共性技术特性的判断，总结出共性技术具有基础性、外部性、集成性和超前性4个特征。

基于指标评估的识别方法通过构建多维度指标体系，对拟识别的技术进行全面、系统地评估和分析，具有一定的科学性，但是鉴于关键共性技术的特征缺乏共识，部分指标的选定依然依赖专家判断，识别结果仍具有争议。本质上，这也是一种主观识别方法。

1.2.3　基于专利网络的识别方法

基于专利网络的识别方法依据专利之间的共现、共类等共性特征，通过量化技术之间的相对距离和技术知识流网络，以专利引文网络和引文链来构建识别框架。该方法早期主要是通过IPC分类号、德温特手工代码、技术共性度指数等来识别共性技术的通用性特性，通过专利被引量、专利申请数量来识别关键技术的重要性[22]。这在一定程度上克服了专家评估的主观性，但受限于IPC分类号或德温特分类代码的分类精度，也忽略了技术知识流的方向性；随后，利用专利之间的引用关系构建专利网络的方法嵌入其中，有效弥补了之前的不足。运用该方法，Feldman等[23]发现关键共性技术具有互补性、适用性和连续性特征；Ho等[24]识别出生物技术领域的关键技术。

基于专利网络的识别方法具有内容客观、标准化程度高的优点，但同时也存在引文时滞性缺点。一项专利从研发、申请、公开到授权，再到专利文献引证，需要漫长的时间，然而，关键共性技术创新是一个动态变化的过程，随着经济和技术的快速发展，其影响因素在不同阶段或者同一阶段不同时期都可能有所差异。因此，该方法识别结果的及时性和有效性有待提升。

1.2.4　基于专利挖掘的识别方法

基于专利挖掘的识别方法通过构建专利文本内容数据集，运用文本聚类、SAO（subject action object）结构、LDA（Latent Dirichlet allocation）主题模型等自然语义处理，挖掘专利文本隐含的技术主题及主题特征词。相较于SAO语义识别技术（难以应对大批量数据、提取文本隐含的技术主题），LDA主题模型通过大幅降低文本维度，从而便于处理海量专利文本数据。许海云等[25]指出主题模型适用于大规模文档隐含主题的深层次挖掘，具有信息挖掘层次较深、语义程度较高的优势。陈伟等[26]分别使用维特比算法（Viterbi algorithm）、LDA主题模型和隐马尔可夫模型对未来技术趋势进行定量预测。马永红等[6]等基于LDA主题模型对新材料专利数据进行研究，发现铝合金制备、纳米粉末及其薄膜制备工艺、金属粉末制备及应用等是新材料领域的关键共性技术。

基于专利挖掘的识别方法通过对海量异质专利文本数据进行建模，按照主体对文本进行聚类，较好克服了上述3种识别方法的客观性不足和时滞等缺陷，具有客观、简便易操作、时间周期短等优势，因此本文选取专利挖掘的识别方法来开展研究。

2　产业关键共性技术识别流程与方法

识别产业关键共性技术的基本思路是针对产业关键共性技术在关键性和共性两个方面的技术性特征，以专利文本为分析对象，采用主题词分析方法，分别挖掘专利技术的关键性、共性特征来识别关键技术和共性技术，并以同时具备这两方面特征的技术作为关键共性技术。为此，首先构建LDA主题模型，包括数据预处理，即运用维特比算法进行分词操作、采用LDA主题模型对文本数据进行主题建模；其次，以主题为节点建立隐马尔可夫链，运用Pagerank算法得到技术主题的关键性得分；再次，从主题的共现矩阵出发，计算技术共现率，得到共性得分；最后，根据关键性得分与共性得分选取合理阈值，以超过阈值的专利主题作为产业关键共性技术主题。如图1所示。

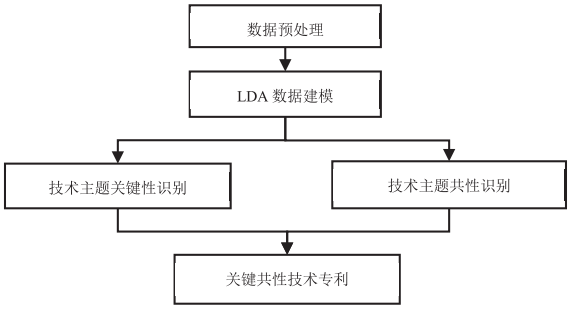

图1　关键共性技术的专利文本识别流程

2.1　构建LDA主题模型

2.1.1　专利摘要文本预处理

本研究以专利文本为对象来识别产业关键共性技术。在对专利文本进行分析、识别专利主题时，首先需要对专利摘要文本进行预处理，即对专利文本进行分词。传统的分词方式往往是采用庞大的词库，综合词库里各个词语的权重来进行操作，但是，专利文本包含了很多专业术语，主流、通用的词库很难将其全部囊括，因而采用传统的分词方式会降低分词的准确性。鉴于此，本研究采用动态规划算法，即通过动态规划寻找概率最大化的词汇切分路径，以识别专利文本中的专业词汇、有效补充分词词库。

具体来说，采用维特比算法来开展分词分析。维特比算法是能够快速求解最佳路径问题的一种动态规划算法，即，在某一时刻 T_n 的最佳路径 S_n 一定满足这一条件：在 T_n 之前的时刻 T_{n-1}，其路径 S_{n-1} 也是最佳的；由此，令上一时刻 T_{n-1} 的最佳路径为 S_{n-1}，将其与上一时刻 T_{n-1} 到现在时刻 T_n 的最短距离 $S_{\min}$ 相加，就得到当前时刻的最佳路径：$S_n=S_{n-1}+S_{\min}$。依此类推，就可以得到全部时刻的最佳路径[27]。运用上述思想，计算概率最大化的词语切分路径，进而补充原有词库中所缺乏的专业术语，能够有效提高分词效果。

2.1.2 LDA 技术主题提取

专利技术主题提取采用潜在狄利克雷分配主题模型算法。LDA 主题模型是一种用于离散数据集合的生成式概率模型，而且是一个三层次（词、主题、文档）贝叶斯概率主题模型，将每一个数据集都视为一组潜在主题的组合，利用主题概率分布对每篇文档进行摘要表达，而且基于词袋模型，在建模过程中忽略单词的词序，化繁为简，为模型的进一步改造提供了条件[28]。具体而言，LDA 算法使用 Dirichle 分布（Dir(·)）刻画文档生成过程。Dirichlet 分布是一种在实数域以正单纯形为支撑集的高维连续概率分布，是 Beta 分布在高维情形的推广和应用，具有共轭性、聚合性、中立性等特点，可以较好地刻画文档生成过程；同时，通过指定主题个数，LDA 算法还能避免模型复杂度过高的问题。

假定专利文本中隐含的技术主题服从以下概率分布：

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

式（1）中：θ_{dk} 表示专利文献 d 在技术主题 k 中的分布；α_k 服从 Dir(β)；$\Gamma(\alpha_k)$ 为 gamma 分布。

对每一个技术主题 k 生成多项分布，对每篇专利文献 d 生成主题词分布 θ_d，假定其服从 Dir(α)；对每篇专利文献的第 n 个主题生成主题项 z_{dn}，令其服从多项式分布$\mathrm{Multinomial}\left(\theta_d\right)$；对每篇专利文献的第 n 个单词生成词项 w_{dn}，令其服从$\mathrm{Multinomial}\left(\phi_{z_{dn}}\right)$。因此，式（1）中 LDA 的似然模型可描述为：

$$p\left(w|\alpha,\beta\right)=\prod_{d=1}^{D}\int p\left(\theta_d|\alpha\right)\prod_{n=1}^{N_d}\Sigma_{Z_d}p\left(Z_{dn}|\theta_d\right)p\left(W_{dn}|\phi_{Z_{dn}}\right)d\theta_d \tag{2}$$

2.1.3 主题数量的选取

运用 LDA 主题模型对文本进行主题的提取与机器学习方法类似，这些方法都需要考虑过拟合的问题，而解决过拟合问题的关键是准确设定目标主题数量，设定准确的主题数量可以提高 LDA 主题模型的精确度。但是，LDA 主题模型自身并不能确定最优的主题数目，目前学界也缺乏统一并且有效的方法。一种广为应用的选取主题数量的参考指标为模型困惑度（perplexity）。在这里，困惑度用来评估语言模型的适用性。其基本逻辑是，被赋予较高概率值的语言模型更适用，并且随着潜在主题数量增加，困惑度呈现递减规律。因而该方法对新文本有较好的预测效果。在 LDA 主题模型中，困惑度计算公式如下：

$$\mathrm{Perplexity}\left(D\right)=\exp\left\{-\frac{\sum_{d=1}^{M}\log p\left(w_d\right)}{\sum_{d=1}^{M}N_d}\right\} \tag{3}$$

式（3）中：D 表示语料库中的测试集；M 为文档篇数；N_d 表示每篇文档 d 中的单词数；w_d 表示文档 d 中的词；$p(w_d)$ 为文档中词语 w_d 产生的概率。

但要指出的是，虽然困惑度能够表征模型的预测能力，但过于强调预测能力则会导致提取的主题数过多。权衡模型的泛化效果和主题抽取能力后，借鉴关鹏等[29]的做法，采用能够兼顾主题间相似度与困惑度的 Perplexity-Var 方法来选择最优主题数量。该方法使用主题间的相对熵（KL 散度）来衡量主题的稳定性，并采用惩罚项的方式来控制过多的主题数量。Perplexity-Var 指标计算公式如下：

$$\mathrm{Perplexity-Var}\left(D_{\mathrm{test}}\right)=\frac{\mathrm{Perplexity}\left(D_{\mathrm{test}}\right)}{\mathrm{Var}\left(D_{\mathrm{test}}\right)} \tag{4}$$

式（4）中：D_{test} 为测试数据集；Perplesity(D_{test}) 为相应的困惑度；Var(D_{test}) 是相应的主题方差，衡量主题之间的稳定性和差异性，利用主题的 JS 散度计算得到。

随后，借鉴 Heinrich[30] 的参数估计方法对部分参数进行设定。将式（2）中的参数分别设定为：$\alpha=\frac{50}{K}$，$\beta=0.1$，进一步使用吉布斯抽样方法得到主题集合，以及每项专利的技术主题归属 $D_k=\{j_1,\cdots,j_n\}$。

2.2 技术主题关键性识别模型（隐马尔可夫过程）

隐马尔可夫过程（Hidden Markov model）是一种动态贝叶斯网络的生成模型，是一种有向图模型，描述了隐马尔可夫链过程生成不可观测的状态随机序列，再由后者生成可观测的随机序列过程，前后生成的两个序列分别称为状态序列和观测序列，序列的位置被视为时刻。隐马尔可夫过程包括 3 个要素：初始概率分布、状态转移概率分布与观测概率分布。假定 Q 为所有可能状态的集合，V 是所有可能的观测状态集合，即：$Q=\{q_1,q_2,\cdots,q_{N'}\}$，

$V=\{v_1,v_2,\cdots,v_{M'}\}$。其中，$N'$是可能的状态数；$M'$是可观测到的数量。同时，设定 I 为长度为 T 的状态序列，O 是对应的观测序列，即$O=\{o_1,o_2,\cdots o_{N'}\}$。

在隐马尔可夫过程，假定状态转移概率矩阵为 $\boldsymbol{A}$，观测概率矩阵为 $\boldsymbol{B}$，初始状态概率向量为 $\boldsymbol{C}$。具体来说，状态转移概率矩阵 $\boldsymbol{A}$ 为：$\boldsymbol{A}=\left[a_{ij}\right]_{N'\times N'}$。其中，$a_{ij}=\left(i_{t+1}=q_j|i_t=q_i\right)$，$i=1,2,\cdots,N'$，$j=1,2,\cdots,N'$，表示的是 t 时刻状态 q_i 在 t+1 时刻变成状态 q_j 的概率。初始状态概率向量 $\boldsymbol{C}$ 为：$\boldsymbol{C}=\left(C_i\right)$。其中，$\boldsymbol{C}_i=P\left(i=q_i\right)$，$i=1,2,\cdots,N'$。$C_i$ 表示的是 t=1 时刻状态为 q_u 的概率，即初始状态的概率。而 $\boldsymbol{B}$ 矩阵为观测序列，即 $\boldsymbol{B}=\left[b_{ij}\right]_{N'\times M'}$。其中，$b_{ij}=p\left(O=j|I=i\right)$，表示观测状态为 j 而可能的状态为 i 的概率。因此一个隐马尔可夫过程可以用三元符号表示为：$\gamma=\{\boldsymbol{A},\boldsymbol{B},\boldsymbol{C}\}$。

2.3　技术主题关键性得分计算方法（PageRank）

技术主题关键性的测度采用 PageRank 算法计算关键性得分。PageRank 算法源于搜索引擎。最早的搜索引擎以人工对网页进行分类，以方便用户分类点击浏览；而后，随着网页海量增加，人工分类已经不能满足用户需求，搜索引擎迈入文本检索时代，但随之而来的是搜索结果质量不高、网页搜索充斥大量垃圾信息，网页排序成为亟待解决的问题。对此，谷歌的创始人谢尔盖·布林[31]针对网页的重要性与关键性问题进行了评价研究，随后提出了用于网页重要性的识别方法，即 PageRank 算法。Pagerank 算法的逻辑如下：一个网页被链接次数越多，表明其 PageRank 值越高；被 PageRank 值高的网页所链接的网页，其 PageRank 值也比较高。就其运用而言，分为两步：首先对每个网页赋予 PR 值；其次通过（投票）的算法进行迭代，达到平稳分布的值即得到了最终得分。PR 值的计算方法主要有幂迭代法、特征值法与代数法。将 PageRank 算法运用于技术主题关键性得分的计算步骤如下：将 LDA 得到的每一个主题作为一个“网页”，依据“网页”链接状态，被“链接”主题的 PageRank 值越高，表明其重要性就越高，相应地，其关键性程度也越高，据此可以计算得到每一个技术主题对应的关键性得分。

2.4　技术主题共性识别模型（技术共现率）

本研究以技术共现率来刻画产业关键共性技术的共性特征。技术共现率为某一技术对其他技术领域的影响，刻画了技术共性特征，基于技术主题共现矩阵计算得到。技术主题共现矩阵是由不同主题之间关键词共同出现的次数所组成的矩阵，即矩阵对角线元素的数值相同，且为矩阵中元素的最大值。技术共现率是共现伙伴数与高频技术领域数的比值，共现伙伴数是与某技术领域存在共现关系的领域数。假设通过 LDA 提取出来的技术主题数一共有 h 个，共现矩阵如表 1 所示，其中 Topic(I") 即为第i"个主题，$Q(i",j")$为第i"个主题与第j"个主题的共现主题词数量。

表 1　技术主题共现矩阵

主题	Topic(1)	Topic(2)	Topic(3)	……	Topic(h)
Topic(1)	Q(1,1)	Q(1,2)	Q(1,3)	……	Q(1,h)
Topic(2)	Q(2,1)	Q(2,2)	Q(2,3)	……	Q(2,h)
Topic(3)	Q(3,1)	Q(3,2)	Q(3,3)	……	Q(3,h)
……	……	……	$Q(i",j")$	$Q(i",i")$	……
Topic(h)	Q(h,1)	Q(h,2)	Q(h,3)	……	Q(h,h)

将阈值设定为 e，则对于主题 i 而言，共现次数大于阈值 e 的主题数量为：

$$\text{num}=\text{count}\left(Q(i",j")>e\right) \quad (5)$$

因而，主题i"的技术共现率为：

$$\text{count}=\frac{\text{num}}{h-1} \quad (6)$$

需要说明的是，对于共现伙伴数量 num，运用能够计算区域（range）中满足给定条件（criteria）的单元格个数的函数，即 Microsoft Excel 软件中的 COUNTIF(range, criteria) 函数来计算。

3　专利数据获取与描述性统计

3.1　专利数据来源与检索方法

3.1.1　专利数据来源

本研究使用的专利数据来源于大为 innojoy 数据库，该数据库为每一条专利提供了完整的数据信息，比如专利文摘、说明书、法律状态、同族专利等，在学界和业界广为应用。由于无法获取科研院所的完整名录，而高校和上市公司名录不难获取，因此根据研究需要和数据可得性，分别选取高等院校和上市公司来进行分析。对于高校，选取大规模合并后的 2001 年作为专利检索起始年份；对于上市公司，选取中国提出自主创新战略的 2006 年作为专利检索起始年份。

3.1.2　专利检索方法

本研究所涉及的专利数据十分庞大，为提高检索效率，采用表达式搜索：AD=（start_year to end_year）and PA=（patent_appliers）。其中，start_year 为专利申请开始年。若检索对象为高校，开始年设定为 2001，若检索对象为上市公司，开始年则设定为 2006。end_year 为专利申请结束年，对高校和上市公司均设定为 2020。patent_appliers 为申请人名单，如果同时有多家单位共同申请，则不同申请人之间用“or”分隔。所选取上市公司范围为 2021 年 6 月

证监会公布的A股全部上市公司（不含被进行退市风险警示的上市公司），共计4 016家；选取的高校范围为教育部的《高等学校科技统计资料汇编》附录中全部本科高校，进一步与近年来更名、合并后的高校名单比对，最终确定的高校数量为690所。

对高校专利而言，假如检索高校为“清华大学”与“北京大学”，其检索式为：AD=（2001 to 2020）and PA=（清华大学 or 北京大学）。对上市公司而言，假如检索公司为“珠海格力电器股份有限公司”与“东风汽车股份有限公司”，检索式为：AD=（2006 to 2020）and PA=（珠海格力电器股份有限公司 or 东风汽车股份有限公司）。按照上述方法，最终获得专利3 216 556件，其中高校专利2 157 526件，上市公司专利1 059 040件。专利总文件大小为4.06 GB，数据十分庞大。

3.2 专利数据的描述性统计

3.2.1 专利总量

首先比较专利申请数量。如图2所示，样本高校申请的专利数量远多于公司的专利数量，前者大约为后者的两倍。原因有多方面：一是高校研发人员数量更多；二是高校和上市公司申请专利的激励不同，高校科研人员在项目申请和结题、职称评审、岗位晋级等诸多方面都需要专利作为支撑，而只有在专利申请量上具有优势才可能获批更多专利，而上市公司在这方面的激励较小，除高新技术企业认定、一些产业支持项目申请需要专利支撑外，除非能够产生显著的经济效应，企业大规模申请专利的动机较小。

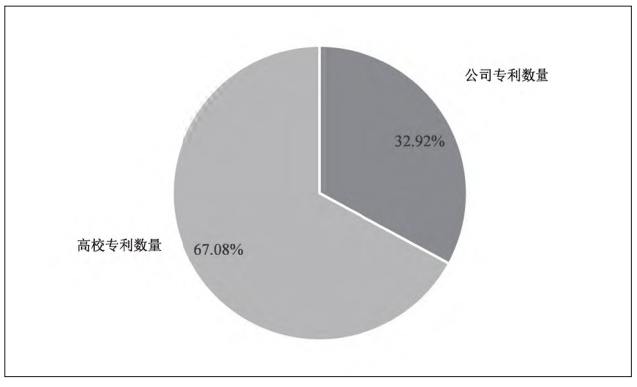

图2 样本高校与上市公司专利申请量占比

样本高校和上市公司的授权专利总数分别为758 413项、606 607项，前者依然高于后者，但与专利申请相比，高校占比下降了大约12个百分点（见图3）。这表明高校申请专利转化为授权专利的比例低于上市公司，即高校申请专利的质量低于上市公司；同时，也意味着高校专利申请具有策略性倾向，即“为申请而申请”，申请专利的创新性、新颖性低于上市公司。换言之，上市公司申请的专利质量更高，更加符合专利授权条件。

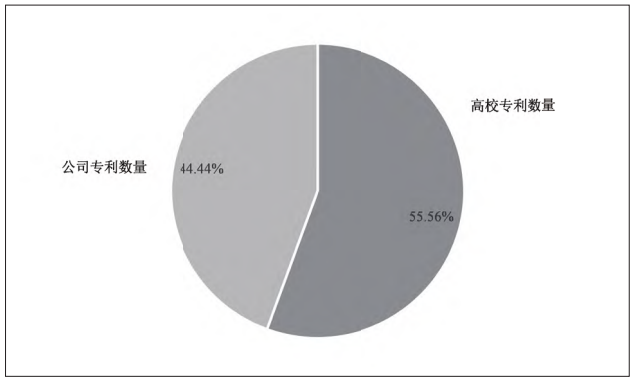

图3 样本高校与上市公司专利授权量占比

3.2.2 专利结构

根据专利的技术含量、授权难易程度不同，中国的专利分为发明、实用新型和外观设计，其中，发明专利技术含量最高，申请难度较大，产业关键共性技术主要分布在这一类专利上。如图4、图5所示，高校和上市公司申请专利的构成有较大的区别：高校专利中发明专利占比高于上市公司专利中发明专利占比；相应地，上市公司申请的实用新型和外观设计专利占比要高于高校。这一差异主要源于两方面：一是各自申请专利的激励不同。发明专利在高校科研项目申请和结题、科研考核中的权重更高、更具市场转让价值；上市公司更看重市场价值，即便是技术新颖度不高的实用新型和外观设计，只要具有一定的市场价值，为占领市场、保持领先，企业也有动机去申请。二是各自的研究禀赋不同。高校重在基础研究和应用研究，研究成果即创新性知识更有助于形成新颖性更高的发明专利；企业重在开展试验与发展，更有助于形成技术新颖度不高的实用新型和外观设计专利。

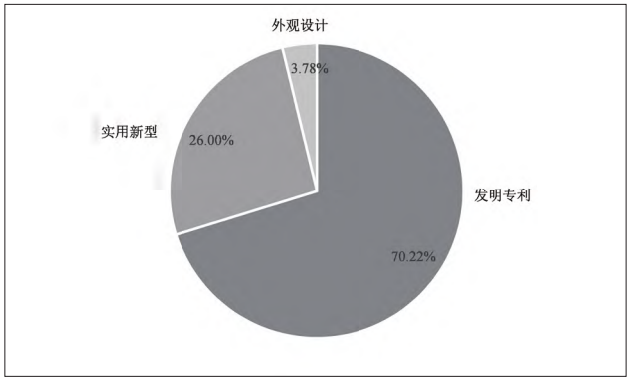

图4 样本高校申请专利结构

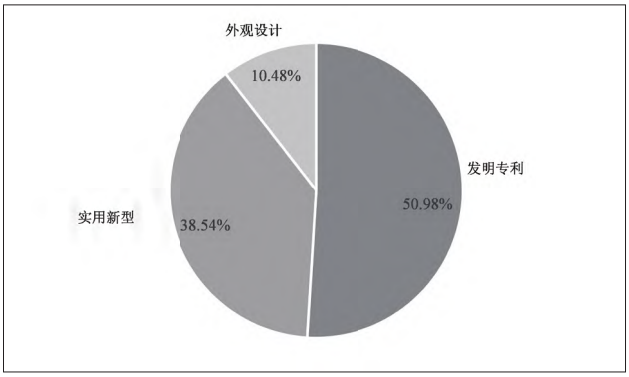

图 5　样本上市公司申请专利结构

进一步分析授权专利构成可以发现（见图 6、图 7），高校和上市公司也呈相同的特征：发明专利占比分别接近七成、五成；相应地，实用新型和外观设计之和占比分别为三成、五成。

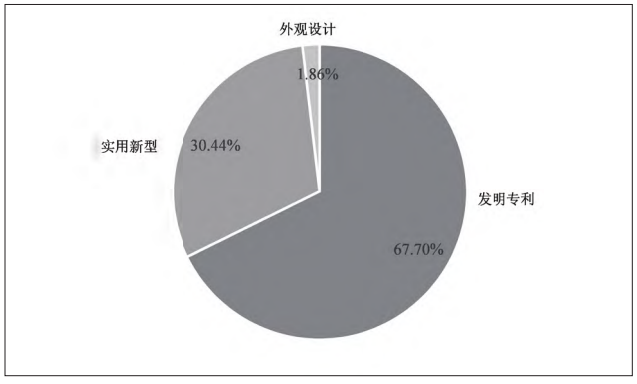

图 6　样本高校授权专利类型占比

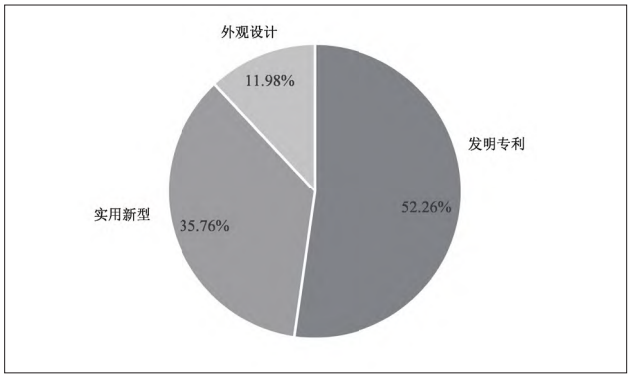

图 7　样本上市公司授权专利类型占比

3.2.3　发明专利总量变化趋势

近年来中国专利呈爆炸式增长，样本高校和上市公司发明专利申请数量增长趋势基本一致（见图 8、图 9）：自 2001 年以来，发明专利申请保持稳定、持续、高速增长；在 2010 年之前，申请专利数量总体较少，增加速度也比较温和；在 2010 年之后，专利数量开始快速增长，尤其是在 2015 年之后，申请专利的增速更快；2020 年由于受新型冠状病毒感染疫情影响，高校和上市公司的发明专利申请均有所下降。

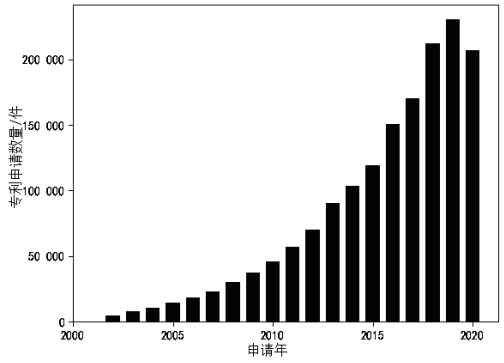

图 8　样本高校发明专利申请趋势

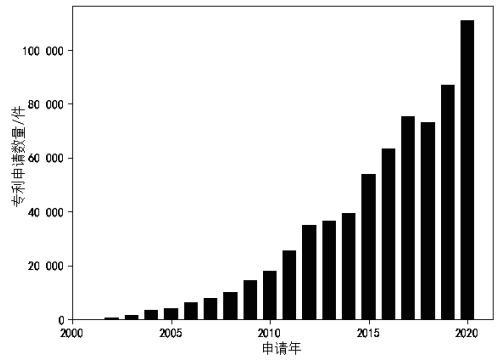

图 9　样本上市公司发明专利申请趋势

从授权专利来看，其趋势与申请专利基本一致（见图 10、图 11）：总体上数量保持增长，但增长具有一定波动。其中与发明专利申请相比，尽管 2020 年发明专利申请数有所下降，但授权发明专利数则比 2019 年大幅增加，这可能与近年来高质量发展背景下国家强调专利质量的专利政策导向有关。

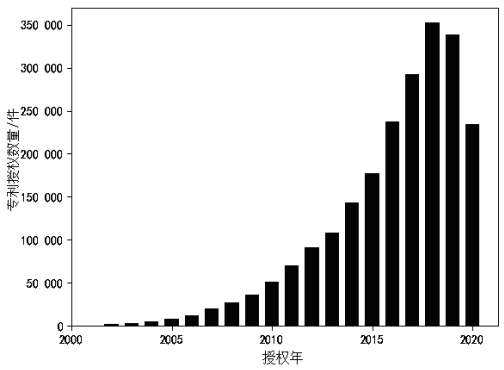

图 10　样本高校发明专利授权趋势

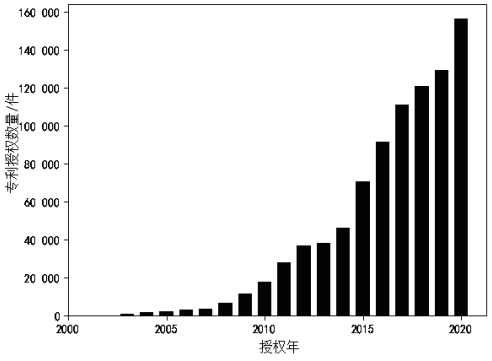

图 11 样本上市公司发明专利授权趋势

3.2.4 发明专利数量较多的高校与上市公司

选取发明专利申请和授权数量最多的样本进行分析，可以发现（见表 2），发明专利申请数量位居前 20 名的高校多为综合性、理工科大学，前 20 名上市公司多为行业领军企业；进一步还可以发现，排名前 20 位高校、上市公司的内部专利申请数量相差极为悬殊，排名第 1 位的高校、上市公司发明专利申请数分别是第 20 位的 2.9 倍、10.1 倍，相对而言，上市公司之间专利申请数量相差更为悬殊。

表 2 申请专利数量前 20 位的样本高校与上市公司

高校名称	数量 / 项	公司名称	数量 / 项
浙江大学	48 830	珠海格力电器股份有限公司	66 766
清华大学	37 181	中兴通讯股份有限公司	53 169
华南理工大学	34 141	中国石油化工股份有限公司	58 120
东南大学	32 671	美的集团股份有限公司	45 501
天津大学	30 360	京东方科技集团股份有限公司	38 472
上海交通大学	28 370	中国石油天然气股份有限公司	29 311
浙江工业大学	26 549	比亚迪股份有限公司	19 065
江南大学	25 968	广东美的制冷设备有限公司	11 911
吉林大学	24 786	九阳股份有限公司	11 345
哈尔滨工业大学	24 746	四川长虹电器股份有限公司	9 674
西安交通大学	21 210	重庆长安汽车股份有限公司	9 523
电子科技大学	20 450	海洋王照明科技股份有限公司	9 070
北京航空航天大学	20 448	北汽福田汽车股份有限公司	9 041
华中科技大学	20 255	宝山钢铁股份有限公司	8 772
昆明理工大学	19 556	长城汽车股份有限公司	8 463
山东大学	19 165	安徽江淮汽车集团股份有限公司	7 606
江苏大学	18 910	鞍钢股份有限公司	7 346
中南大学	17 887	广州视源电子科技股份有限公司	7 086
北京工业大学	17 386	潍柴动力股份有限公司	6 684
山东科技大学	16 734	高通股份有限公司	6 482

分析授权专利数量位居前 20 位的样本高校与上市公司可以发现（见表 3），上市公司的授权专利数量更集中于头部公司，而高校则相对平均；同时，与专利申请前列名单较为一致，排名前 20 的高校中多是“985 工程”/“211 工程”高校或者是工科强校，而上市公司中，制造业企业占比最高，其中中国石油化工股份有限公司及其子公司所占比例最高。此外还可以发现，专利数量排名前 20 位的高校与上市公司的授权比例均不高。

表 3 授权专利数量前 20 位的样本高校与上市公司

高校名称	数量 / 项	公司名称	数量 / 项
清华大学	17 130	中国石油化工股份有限公司	41 293
浙江大学	16 008	中兴通讯股份有限公司	16 300
哈尔滨工业大学	10 933	京东方科技集团股份有限公司	15 348
东南大学	10 373	珠海格力电器股份有限公司	11 401
华南理工大学	10 202	中国石油天然气股份有限公司	8 076
西安交通大学	9 027	美的集团股份有限公司	7 781
华中科技大学	9 005	比亚迪股份有限公司	5 695
上海交通大学	8 839	宝山钢铁股份有限公司	3 022
天津大学	8 700	中联重科股份有限公司	2 183
北京航空航天大学	8 588	海洋王照明科技股份有限公司	1 948
电子科技大学	7 660	广东美的制冷设备有限公司	1 922
西安电子科技大学	7 317	北汽福田汽车股份有限公司	1 754
中南大学	7 194	安徽江淮汽车集团股份有限公司	1 752
浙江工业大学	7 074	鞍钢股份有限公司	1 750
山东大学	6 909	四川长虹电器股份有限公司	1 667
江南大学	6 114	北京京东方显示技术有限公司	1 637
大连理工大学	6 050	武汉华星光电技术有限公司	1 554
北京工业大学	6 025	北京京东尚科信息技术有限公司	1 326
南京航空航天大学	5 595	上海华力微电子有限公司	1 221
北京理工大学	5 507	烽火通信科技股份有限公司	1 056

4 计算结果

4.1 专利技术的关键性、共性得分

在计算专利技术的关键共性得分时，其参数与阈值（包含最优状态转移概率矩阵的概率阈值、共性矩阵的频次阈值）选择上参考陈伟等[26]提出的“双百”原则，即将概率阈值设定为 1%，频次阈值设定为 100。具体来说，对于隐马尔可夫过程中的参数，初始状态概率矩阵依照同等未知原则，即初始状态在没有给出较多信息的时候，将不同主题的初始状态视为相同地位，从而选取均匀分布，得到最优状态转移概率矩阵。将概率阈值设定为 1%，共性矩阵的频次阈值设定为 100。在进行关键性 PageRank 计算时，阻尼因子与收敛阈值的选取只要达到计算收敛要求即可。经过多次尝试，将转移的阻尼因子设定为通用性较强的 0.85，收敛阈值设定为 0.000 01，经计算后发现满足收敛要求，这表明收敛阈值的选取是合理的。

4.1.1 高校

根据 LDA 模型筛选出的技术主题，分别计算其关键性得分和共性得分。如图 12 所示，关键性得分位于 X 轴，共性得分位于 Y 轴，沿 X 轴越往右延伸表示技术主题的关键性程度越高，沿 Y 轴越往上延伸表示技术主题的共性程度越高。X 轴的最小值为 0.000 5，这些值对应的技术主题可以归类为边缘计算主题，相当于共性水平中的 0 值，即下确界。从图 12 来看，散点主要分布在平行于横轴和纵轴的两

条线带上。从平行于纵轴的线带来看，关键性得分处于最低值即 0.000 5 处的技术主题非常多（位于该线带上的散点远多于其他散点），表明高校专利技术中属于关键性主题的技术很少，多为非关键性技术主题；换言之，高校专利中具有重大关键性、突破“卡脖子”技术的专利数量还比较有限，提高专利技术的关键性能仍然任重道远。从平行于横轴的线带来看，共性得分较低的散点分布也较为集中，即高校专利技术中共性程度低的技术主题较多；换言之，专有技术主题的专利较多。除这两条线带以外的点，即兼具关键性与共性特征的专利技术分布较为稀疏、零散，尤其是位于图 12 右上角的区域，即关键性与共性技术程度都比较高的点很少，表明高校专利中属于关键共性技术的专利比较少。从总体上来看，高校申请专利所对应的技术主题相对离散，关键性低或共性度低的专利是主流，能同时兼顾关键性与共性的技术主题很少，大量技术偏离关键共性技术主题，处于相对边缘地位。

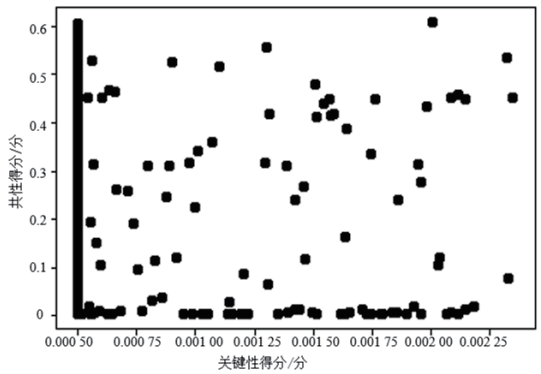

图 12　样本高校专利关键性和共性得分散点分布

4.1.2　上市公司

如图 13 所示，上市公司专利技术的关键性、共性的分散点主要分布在位于左侧的一条线带上，即大量散点分布在关键性得分最低值水平（0.000 5）的垂线上，表明企业专利技术主题中的关键性技术非常少，多为非关键技术；同时，右边区域的散点相对较为集中（但数量远低于左侧线带上的散点），即在上市公司专利技术主题中，关键性较高而共性度较低和关键性较高且共性度较高的技术主题并存。与图 12 相比，图 13 中位于右边区域的散点相对较多、中间区域散点较少，即上市公司专利中属于关键性技术主题的比例略高于高校。进一步还可以发现，图 13 中位于右上方区域的散点较为集中，即上市公司申请的专利技术中关键性与共性得分都比较高的技术主题数多于高校。

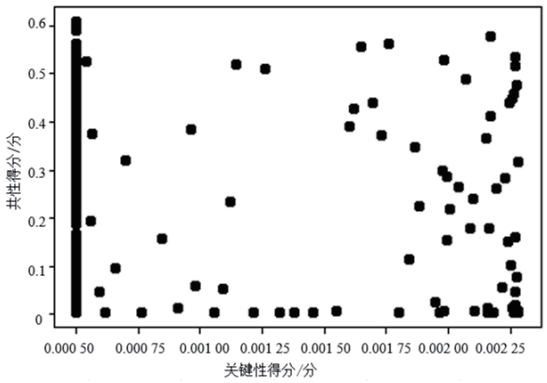

图 13　样本上市公司专利关键性和共性得分散点分布

4.2　阈值选择

对产业关键共性技术主题的识别需要兼顾关键性得分与共性得分两个维度，为此，需要分别选择二者的阈值，只有当技术主题的关键性得分、共性得分均超过所设定的阈值时，该技术主题才能被认定为关键共性技术主题。因此，从专利文本中识别出来的关键共性技术主题与所设定的阈值密切相关：关键性得分阈值设定越高，则所识别出来的主题更具备关键性属性；共性得分阈值设定越高，则所识别出来的主题更具共性属性；两个指标阈值均设定越高，则所识别出来的主题同时具有更高的关键性和共性属性。

阈值选择以我国《产业关键共性技术发展指南（2017 年）》（以下简称《指南》）为关键共性技术主题标准，通过分析阈值选取后得到的关键共性技术主题与《指南》的相似度，来识别高校和上市公司专利中有哪些是属于关键共性技术。假如设定的关键性阈值为 a、共性度阈值为 b、相似度临界值为 c，如果基于 a、b 识别出来的技术主题与《指南》的相似度大于或等于临界值 c，则相关技术主题就属于关键共性技术；如果识别出来的技术主题与《指南》的相似度低于临界值 c，则需要进行阈值的调整。此时可以将两个阈值（a、b）一个调大一个调小，或者一个调小一个调大，最终使相关系数达到临界值。具体调整的过程如图 14 所示。也就是说，在实际操作过程中，可以选取固定数值的间隔，通过不断试错得到一个解，注意这里并非唯一解。

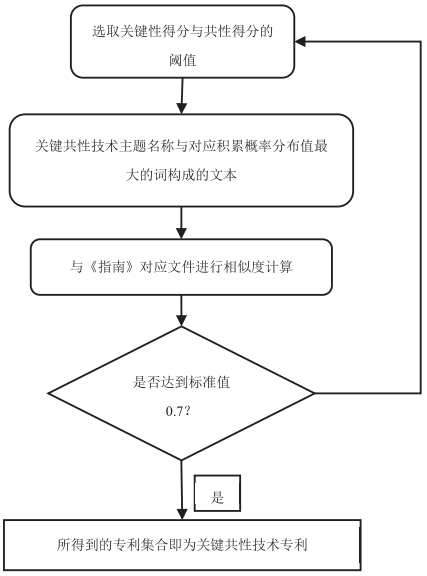

图 14 关键共性技术阈值调整流程

需要指出的是，阈值选择的关键是计算文本相似度。近年来，文本相似度分析方法广为应用，比如，姜富伟等[32]运用文本相似度来计算货币政策文本的情感倾向，以此来研究其与股票市场价格反应的关系；覃飞等[33]采用文本相似度划分了企业年报是否与当地政府产业政策具有关联性。文本相似度是指两个或多个实体（词语、短文本、文档）之间的相似程度，也可以理解为文本之间的共性或差异性，共性越大、差异越小，相似度就越高；当两个文本片段完全相同时，相似度最大。本研究采用具有较高稳健性的百度人工智能（AI）开发平台（以下简称“百度平台”）中长文本相似度的应用程序接口（API）。百度平台具有强大的词库存储，相较于个人训练模型，稳健性更高。具体来说，阈值选择方法如下：

首先，将识别出来的产业关键共性技术主题对应的 LDA 分词组合形成关键共性技术主题文本，运用余弦相似度方法将其与《指南》进行文本相似度分析，得到两篇文本之间的相关系数（采用 2015、2013、2011 年度《指南》估计的结果相近）。根据广为应用的高相关性门槛值标准即 0.7[34]，只要测度出的相关系数大于 0.7，就可以将识别出的主题认定为产业关键共性技术主题；相关系数小于 0.7 的，则进行阈值调整。

其次，进一步检验前述方法的合理性。将剩余的主题，即从属于非产业关键共性技术主题的 LDA 分词组成另一篇文本——非关键共性技术主题文本。将两篇文本分别与《指南》进行相似度测算，得到各自的相关系数，分别为 e_1 和 e_2。如果 e_1 大于 e_2，则表明产业关键共性技术主题识别中的阈值选择是合理的；如果 e_1 小于或等于 e_2，则需要进行阈值的调整。阈值的调整方法同上。

4.3 识别结果

匹配结果发现：样本高校申请的专利中，属于关键共性技术的有 316 948 项，占比为 14.7%（专利申请总量为 2 157 526 件）；上市公司申请的专利中，属于关键共性技术的有 108 494 件，占比为 10.2%（专利申请总量为 1 059 040 件）。上述二者占比的均值为 13.23%，其中高校申请、属于关键共性技术的专利占比略高于上市公司。要指出的是，专利的关键得分和共性得分的组合中，并非只有“高 – 高”组合才属于关键共性技术专利，只要符合阈值条件，即便是“高 – 低”“低 – 高”组合也属于关键共性技术。同时可见，对授权专利而言，高校授权专利中属于产业关键共性技术的为 104 753 件，占比为 13.8%（授权专利总数为 758 413 件）；上市公司专利中属于产业关键共性技术的专利数为 61 372 件，占比 10.1%（授权专利总数 606 607 件）。上述二者占比的均值为 12.17%，其中高校授权专利中属于产业关键共性技术的比例也略高于上市公司。

进一步分析样本高校和上市公司的典型关键共性技术，比较图 12 和图 13 可以发现，有很大一部分主题缺失关键性或共性。为了更好地展示识别结果，在这里选取 X 轴与 Y 轴上最小值之外的第一个刻度值作为阈值展示。为使选取的刻度值清晰，将关键性得分阈值设定为 0.000 75，共性得分阈值设定为 0.100 00，得到高校识别出的关键共性技术主题为 54 个，上市公司识别出的关键共性技术主题为 39 个。将累积概率分布值位居前3位的词语作为主题名称，样本高校与上市公司关键共性得分位居前 15 位的技术主题如表 4 所示，其中“电路芯片耦合”“能效模型图像”“系统数据模块”等与集成电路、人工智能、大数据等近年来广为关注的关键共性技术非常吻合，尤其是集成电路（芯片）是我国面临的“卡脖子”技术难题。这一分析表明，本研究识别出来的技术主题具有关键性和共性属性，具备属于关键共性技术的合理性。

表 4 样本高校与上市公司前 15 位关键共性技术主题

项目	高校	企业
主题	模型路灯图像；涡流模块控制；信号相位信息；光催化剂制备可见光；雨水储液罐体；超声波腐蚀疲劳；能效模型图像；数据浮选杯体；抗体检测抗原；模块控制采集；模型预测查询；钢管输电线路；系统数据模块；挂钩系统风管；抗生素六边形植被	感应锅具表面；单元进水冷凝水；虚拟认证模块；电路芯片耦合；压板支撑杆取样；信息列表声音；插件复位内筒；节目遥控器变化；活动固定窗口；灌装预制锥形；组合制备重量；出现脂肪酸拆分；天线碰撞信号；单元系统扫描；检测电路定时

5　结论

本研究运用专利挖掘方法，通过构建 LDA 主题模型、计算技术主题的关键性得分和共性得分、选择得分阈值等方法，通过我国高校和上市公司专利来识别关键共性技术。对 2002—2020 年高校专利和 2006—2020 年上市公司专利的描述性统计和技术分析发现：第一，高校与上市公司专利在数量与结构上存在显著差异。从专利数量来看，高校数量远高于上市公司数量，但高校授权专利比例低于上市公司，表明高校申请专利的质量有待提升；从专利结构来看，高校专利多为发明专利，而上市公司专利更偏向实用新型和外观设计专利；另外，从发明专利整体变化趋势来看，高校与上市公司发明专利申请与授权数量均保持稳定、持续、高速增长的趋势。第二，从关键共性技术识别结果来看，无论是高校还是上市公司，属于关键共性技术专利的比例均不高。具体来看，高校、上市公司的申请专利中属于关键共性技术的比例分别为 14.7%、10.2%，授权专利中属于关键共性技术的比例分别为 13.8%、10.1%，高校关键共性技术专利占比略高于上市公司；同时还发现，识别出来的技术主题名称与当前广受关注的关键共性技术密切相关。

研究结果与现实吻合初步验证了本研究方法的可行性和有效性，但本研究也存在以下不足之处：（1）在进行 LDA 技术主题建模过程中，忽略了单词的词序，使得所提取的关键词之间缺少连贯性，文档表征能力有待增强；（2）innojoy 数据库中的专利数据仅公布摘要文本，未包含论文数据，数据覆盖面有待提升。在未来的研究中，将进一步提高技术主题识别的准确性，以及结合专利、论文、研究报告等数据以实现产业关键共性技术的高效预测。

参考文献：

[1] 江鸿，石云鸣．共性技术创新的关键障碍及其应对：基于创新链的分析框架［J］．经济与管理研究，2019, 40(5):74–84.
[2] 刘凤朝，马荣康，孙玉涛．基于专利技术共现网络的纳米技术演化路径研究［J］．科学学研究，2012, 30(10):1500–1508.
[3] 栾春娟．战略性新兴产业共性技术测度指标研究［J］．科学学与科学技术管理，2012,33(2):11–16.
[4] 朱桂龙，朱明晶，尹潇．知识扩散视角下共性技术的商业化评价：基于多层网络的反向识别方法［J］．科学学与科学技术管理，2019,40(2):16–25.
[5] 陈伟，林超然，孔令凯，等．基于专利文献挖掘的关键共性技术识别研究［J］．情报理论与实践，2020, 43(2): 92–99.
[6] 马永红，孔令凯，林超然，等．基于专利挖掘的关键共性技术识别研究［J］．情报学报，2020,39(10): 1093–1103.
[7] GRANBERG A, STANKIEWICZ R. The development of "generic technologies": the cognitive aspects［M］. Lund: Research Policy Institute, 1981:196–225.
[8] TASSEY G. Choosing government R&D policies: tax incentives vs. direct funding［J］. Review of Industrial Organization, 1997,11(5):579–600.
[9] 李纪珍．产业共性技术：概念、分类与制度供给［J］．中国科技论坛，2006(3):45–47,55.
[10] 马名杰．政府支持共性技术研究的一般规律与组织［J］．中国制造业信息化，2005,34(7):14–16.
[11] 周国华，谭晶菁．复杂产品关键共性技术供给模式比较研究［J］．软科学，2018,32(6):97–102.
[12] KEENAN M. Identifying emerging generic technologies at the national level: the UK experience［J］. Journal of Forecasting, 2003,22(2/3):129–160.
[13] 袁思达．技术预见德尔菲调查中共性技术课题识别研究［J］．科学学与科学技术管理，2009,30(10): 21–26.
[14] 曹旭华，王富忠，胡华敏．浙江省新材料行业关键共性技术的技术预见研究［J］．科技进步与对策，2010,27(16):46–49.
[15] 张乔木．基于德尔菲调查和聚类分析的关键共性技术预见研究：以山西省新材料行业为例［J］．科技管理研究，2017,37(13):121–124.
[16] 王秀红，高敏．基于 BERT–LDA 的关键技术识别方法及其实证研究：以农业机器人为例[J].图书情报工作，2021,65(22):114–125.
[17] 黄鲁成，张静．基于专利分析的产业共性技术识别方法研究［J］．科学学与科学技术管理，2014,35(4): 80–86.
[18] LEE C, KWON O, KIM M, et al. Early identification of emerging technologies: a machine learning approach using multiple patent indicators［J］.Technological Forecasting and Social Change, 2018, 127(2):291–303.
[19] 王燕鹏，韩涛，陈芳．融合文献知识聚类和复杂网络的关键技术识别方法研究［J］．图书情报工作，2020,64(16):105–113.
[20] LIU X, PORTER A L. A 3–dimensional analysis for evaluating technology emergence indicators［J］. Scientometrics, 2020,124:27–55.
[21] 江娴，魏凤．基于专利分析的共性技术识别研究框架［J］．情报杂志，2015,34(12):79–84.
[22] MOSER P, NICHOLAS T. Was electricity a general purpose technology? Evidence from historical patent citations［J］. American Economic Review, 2004,94(2):388–394.
[23] FELDMAN M P, YOON J W. An empirical test for general purpose technology: an examination of the Cohen–Boyer rDNA technology［J］.Industrial and Corporate Change, 2012,21(2):249–275.
[24] HO M H–C, LIN V H, LIU J S. Exploring knowledge diffusion among nations: a study of core technologies in fuel cells［J］. Scientometrics, 2014,100(1):149–171.
[25] 许海云，王振蒙，胡正银，等．利用专利文本分析识别技术主题的关键技术研究综述[J].情报理论与实践,2016,39(11):131–137.
[26] 陈伟，林超然，李金秋，等．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［J］．情报学报，2018,37(7):732–741.
[27] 邹佳伦，文汉云，王同喜．基于统计的中文分词算法研究［J］．电脑知识与技术，2019,15(4):149–150,153.
[28] 韩亚楠，刘建伟，罗雄麟．概率主题模型综述［J］．计算机学报，2021,44(6):1095–1139.
[29] 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，2016,32(9):42–50.
[30] HEINRICH G. Parameter estimation for text analysis［R］. Darmstadt: Fraunhofer IGD, 2005.
[31] BRIN S, PAGE L. The anatomy of a large–scale hypertextual web search engine［J］.Computer Networks and 综合业务数字网 Systems, 1998,30(1/7):107–117.
[32] 姜富伟，胡逸驰，黄楠．央行货币政策报告文本信息、宏观经济与股票市场［J］．金融研究，2021,42(6): 95–113.
[33] 覃飞，沈艳．产业政策关联度对公司业绩影响研究［J］．数量经济技术经济研究，2021,38(9):117–138.
[34] 陈希孺．概率论与数理统计［M］．合肥：中国科学技术大学出版社，2009:154–156.

作者简介：胡凯（1975—），男，湖北天门人，硕士生导师，教授，博士，主要研究方向为技术创新经济学；谢芬（1997—），女，湖北恩施人，硕士研究生，主要研究方向为数字金融；杨滨瑜（1998—），男，江苏徐州人，硕士研究生，主要研究方向为产业经济学；胡新杰（1999—），男，江苏徐州人，硕士研究生，主要研究方向为金融科技；刘汉霞（1977—），通信作者，女，湖北麻城人，副教授，博士，主要研究方向为教育管理学。