Statistics and Application 统计学与应用, 2022, 11(1), 169-178
Published Online February 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.111019

Hans汉斯

基于文本挖掘技术的人工智能领域研究热点分析

孟晓宇

燕山大学理学院，河北 秦皇岛

收稿日期：2022年1月21日；录用日期：2022年2月6日；发布日期：2022年2月21日

摘 要

在数字经济不断推进的大背景下，人工智能发展迅速并与多种应用场景深度融合，精准把握人工智能领域的近期发展态势和热点方向具有重要意义。通过使用网络爬虫技术获取人工智能领域高水平文献信息，采用文本挖掘技术，对文献关键词进行共词分析，构建基于摘要的LDA主题发现模型，利用Gephi和Jupyter Notebook软件绘制人工智能领域的知识图谱。分析发现，近三年我国人工智能领域研究最热点关键词为：大数据、深度学习、机器学习；最热研究方向为：高新技术融合、智慧教育、智能医疗、担忧与决策。

关键词

共词分析，LDA主题模型，知识图谱，人工智能

Analysis of Research Hotspots in the Field of Artificial Intelligence Based on Text Mining Technology

Xiaoyu Meng

School of Science, Yanshan University, Qinhuangdao Hebei

Received: Jan. 21st, 2022; accepted: Feb. 6th, 2022; published: Feb. 21st, 2022

Abstract

Under the background of the continuous promotion of digital economy, AI develops rapidly and is deeply integrated with a variety of application scenarios. It is of great significance to accurately grasp the recent development trend and hot direction in the field of AI. By using web crawler

文章引用: 孟晓宇. 基于文本挖掘技术的人工智能领域研究热点分析[J]. 统计学与应用, 2022, 11(1): 169-178.
DOI: 10.12677/sa.2022.111019

technology to obtain high-level literature information in the field of AI, using text mining technology to analyze the co-words of literature keywords, construct LDA topic discovery model based on abstract, and draw the knowledge map in the field of AI by using Gephi and Jupyter Notebook software. It is found that the hottest keywords in the field of AI in China in recent three years are: big data, deep learning and machine learning; the hottest research directions are: high-tech integration, smart education, smart medicine, worry and decision-making.

Keywords

Co-Word Analysis, LDA Topic Model, Knowledge Graph, Artificial Intelligence

1. 引言

为抢抓人工智能发展的重大战略机遇，构筑我国人工智能发展的先发优势，近年来我国在人工智能领域密集出台相关法律法规及政策。例如，2017 年 7 月，国务院发布《新一代人工智能发展规划的通知》，描绘了未来十几年我国人工智能发展的宏伟蓝图，确立了“三步走”目标。2020 年 7 月，五部门印发《国家新一代人工智能标准体系建设指南》，明确提出建立国家新一代人工智能标准体系，加强标准顶层设计与宏观指导。世界各国也均加快了人工智能方面的研发和产业布局，发展人工智能相关理论、算法和技术已经成为当前各主要经济体在未来保持国际竞争和技术优势的重要手段，并已经成为各国科技或产业政策的重要内容。

在上述背景下，近年来人工智能相关研究呈现快速增长趋势，相关研究文献也出现了高速增长。然而文献的种类繁杂、质量参差不齐，不同种类之间又存在信息重复等各种各样的问题，因而在短时间内检索到高质量的满足自己需求的人工智能领域文献以及从检索到的文献中快速发现更多隐含的信息，是研究者面临的一大难题。

为对人工智能的发展和热点进行精准把握，近年来学者们进行了丰富的研究，主要包括相关规划、政策和专利分析，但对人工智能领域的研究文献进行分析的较少。如臧维等[1]，袁野等[2]，李首骏[3]，高长元等[4]分别对我国人工智能相关的政策和合作专利进行文本量化研究，探索前沿趋势。李牧南和王雯殊[5]基于主题建模的文本分析思路进行内容挖掘，研究人工智能相关科学主题的演进模式。吕一博等[6]，魏雪飞[7]采用文献分析软件(如 CiteSpace 和 Endnote)，按照内置文献计量方法进行可视化分析。在数据的选取上，大多学者选择研究人工智能在某一领域的热点内容，如医疗、教育等方面[7] [8]，对人工智能主题文献的综合研究较为匮乏。在上述研究基础上，本文将中文期刊论文、学位论文以及外文期刊的中国作者论文作为数据来源，将近三年人工智能有关的高水平文献作为研究对象，采用共词分析法、构建 LDA 主题发现模型，对文献的关键词和摘要进行两方面研究，绘制人工智能领域知识图谱，可更为全面的揭示研究热点，追踪人工智能技术的发展以及在其他领域应用的前沿问题。

2. 模型理论及建立

2.1. 共词分析理论

共词分析本质上是一种共现分析方法，基本原则是先统计一组关键词中任意两个词在一组文档里某

一篇中是否共同出现，再统计这种共同出现情况的次数并构建对称关键词共现矩阵。在共现矩阵中关键词共现次数可能相差较大不利于数据分析，需要把关键词共现矩阵进行归一化处理转化为关键词相关矩阵。其中，文档-关键词矩阵中“1”表示某个关键词在某篇文档中至少出现过一次，即这个关键词属于这篇文档；共现矩阵中对角线数字表示该关键词在一组文档中的词频，其余位置数字表示任意两个不同关键词在一组文档中共同出现过文档的篇数。

2.2. LDA 主题模型的建立

话题模型主要用于处理离散型数据如文本集合，在信息检索、自然语言处理等领域又广泛的应用。其中隐狄利克雷模型(LDA)是话题模型的典型代表。假定数据集包含 K 个话题和 T 篇文档，用 T 个 N 维向量 $W=\{w_1,w_2,\cdots,w_T\}$ 表示数据集即文档集合，K 个 N 维向量 $(k=1,2,\cdots,K)$ 表示话题，$w_t\in R^N$ 的第 n 个分量 $w_{t,n}$ 表示文档 t 中词 n 的词频，$\beta_k\in R^N$ 的第 n 个分量 $\beta_{k,n}$ 表示话题 k 中词 n 的词频。通过统计文档中出现的词来获得词频向量 $w_i\,(i=1,2,\cdots,T)$，LDA 认为每篇文档包含多个话题，用向量 $\theta_t\in R^K$ 表示文档 t 中所包含的每个话题 k 的比例，进而通过以下步骤由话题生成文档 t：

1) 根据参数为 α 的狄利克雷分布随机采样一个话题分布 θ_t；

2) 按如下步骤生成文档中的 N 个词：

a) 根据 θ_t 进行话题指派，得到文档 t 中词 n 的话题 $z_{t,n}$；

b) 根据指派的话题所对应的词频分布 β_k 随机采样生成词。

于是，LDA 模型对应的概率分布为

$$p(W,z,\beta,\theta|\alpha,\eta)=\prod_{t=1}^{T}p(\theta_t|\alpha)\prod_{i=1}^{K}p(\beta_k|\eta)\left(\prod_{n=1}^{N}p(w_{t,n}|z_{t,n},\beta_k)\,p(z_{t,n}|\theta_t)\right)$$

其中 $p(\theta_t|\alpha)$ 和 $p(\beta_k|\eta)$ 通常分别设置为以 α 和 η 为参数的 K 维和 N 维狄利克雷分布，如

$$p(\theta_t|\alpha)=\frac{\Gamma\left(\sum_k\alpha_k\right)}{\prod_k\Gamma(\alpha_k)}\prod_k\theta_{t,k}^{\alpha_k-1}$$

其中 $\Gamma(\cdot)$ 是 Gamma 函数。显然，α 和 η 是式中待确定的参数。

给定训练数据 $W=\{w_1,w_2,\cdots,w_T\}$，LDA 的模型参数需通过极大似然法估计，即寻找 α 和 η 以最大化对数似然

$$L(\alpha,\eta)=\sum_{t=1}^{T}\ln p(w_t|\alpha,\eta)$$

但由于 $p(w_t|\alpha,\eta)$ 不易计算，上式难以直接求解，因此实践中常采用变分法来取近似解。若模型已知，即参数 α 和 η 已确定，则根据词频 $w_{t,n}$ 来推断文档集所对应的话题结构(即推断 θ_t，β_k 和 $z_{t,n}$)可通过求解

$$p(z,\beta,\theta|W,\alpha,\eta)=\frac{p(W,z,\beta,\theta|\alpha,\eta)}{p(W|\alpha,\eta)}$$

然而由于分母上的 $p(W|\alpha,\eta)$ 难以获取，上式难以直接求解，因此本文采用变分法进行近似推断。

3. 实证分析

3.1. 数据获取及预处理

数据库通常具有下载文献相关信息的功能，但下载数量有限制，如果研究者所需文献数量较大，文

献数据获取将会损失较多科研时间。识别数据库反爬虫机制、使用 Python 软件编写爬虫程序，可快速获取所需信息。

本文使用 Request 模块抓取网页数据，Lxml 模块解析数据，爬取 2019~2021 年以人工智能为关键词的高水平文献，共计 10315 篇。数据来源于中国知网(CNKI)数据库和 Web Of Science 数据库。高水平文献含三部分：第一部分是中文期刊文献。中文期刊需被北大核心、SCI、EI 或 CSCD 收录，这类期刊具有信息量大、利用率高、权威性强等特点，数据涉及文献篇名、作者、刊名、发表时间、关键词、摘要、机构、文章链接等。第二部分为人工智能领域被 CNKI 收录的硕博士学位论文，硕士和博士是科研的中坚力量，他们的研究课题一定程度上代表着我国人工智能领域的科研水平。第三部分是被 Web of Science 收录的中国学者发表的高水平人工智能文献。其中，外文期刊需被 SCI 收录，SCI 收录的文献能够全面覆盖全世界最重要和最有影响力的研究成果。具体检索时，作者关键词选择为 Artificial Intelligence，地区选择为 CHINA，文献类型仅限于论文，并排除综述类文章，确保后续分析的准确性。爬取得到中文期刊文献 5575 篇，硕博学位论文 2305 篇，中国作者发表在外文期刊的文献 2255 篇。

爬取到的文献数据参差不齐，需对数据进行预处理，便于后续分析。首先，文献根据研究需要需删除研讨会综述、课题介绍、会议通知、卷首语、会议记录、课题通过鉴定、读后感、简介、研讨会简介、书评、成果鉴定会、学院以及学校简介信息、人物专访、投稿须知、会议纪要、出版信息、目录信息、公告等无关内容。其次，进行基础预处理操作：删除摘要或者关键词缺失的文献、删除重复项。得到有效中文期刊文献 4896 篇，学位论文 1828 篇，中文作者在国外期刊发表的论文 1715 篇，共计 8439 篇。最后，进行进一步的预处理操作，对英文大小写进行统一化处理、对文献摘要进行分词处理、去除停用词，对文献关键词进行分列。预处理流程图如图 1 所示。

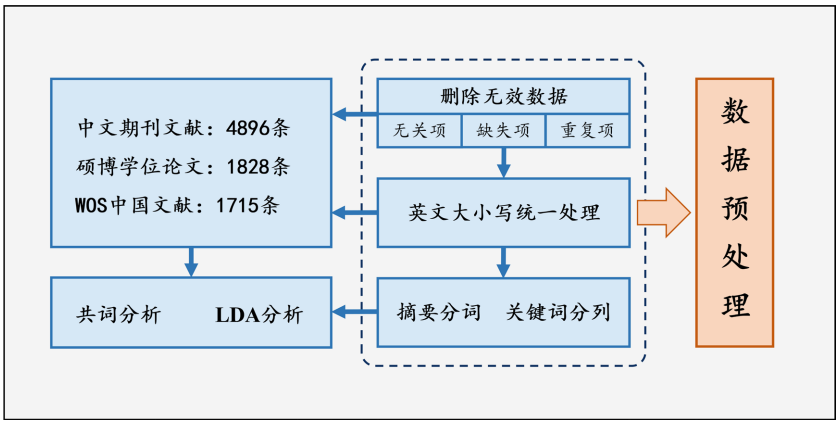

Figure 1. Pretreatment flow chart
图 1. 预处理流程图

3.2. 文献基本信息描述

本文通过运用文献的描述计量方法，以在 CNKI 和 Web of Science 中爬取到的近三年人工智能领域的有效文献数据为基础，展示文献的来源单位、外文期刊收录情况、文献作者群等，反映人工智能领域的研究现状。

研究发现北京邮电大学、吉林大学、电子科技大学等院校对于人工智能领域的相关研究较多，近三年北京邮电大学相关论文量可达 60 篇，硕博学位论文排名前 20 名的授予单位见图 2。在外文期刊中，中文作者发表在 IEEE Access 期刊的该领域论文较多，近三年相关文献达 253 篇，Neurocomputing 和 IEEE

Network 期刊相关论文数量位居二、三位，均在 30 篇左右；排名前 20 名的外文期刊(见图 3)中近 50%的期刊由 IEEE 出版。通过统计文献作者的发文频次，研究发现中文期刊文献中肖锋、何大安等学者在该领域的发文量较多，见图 4。

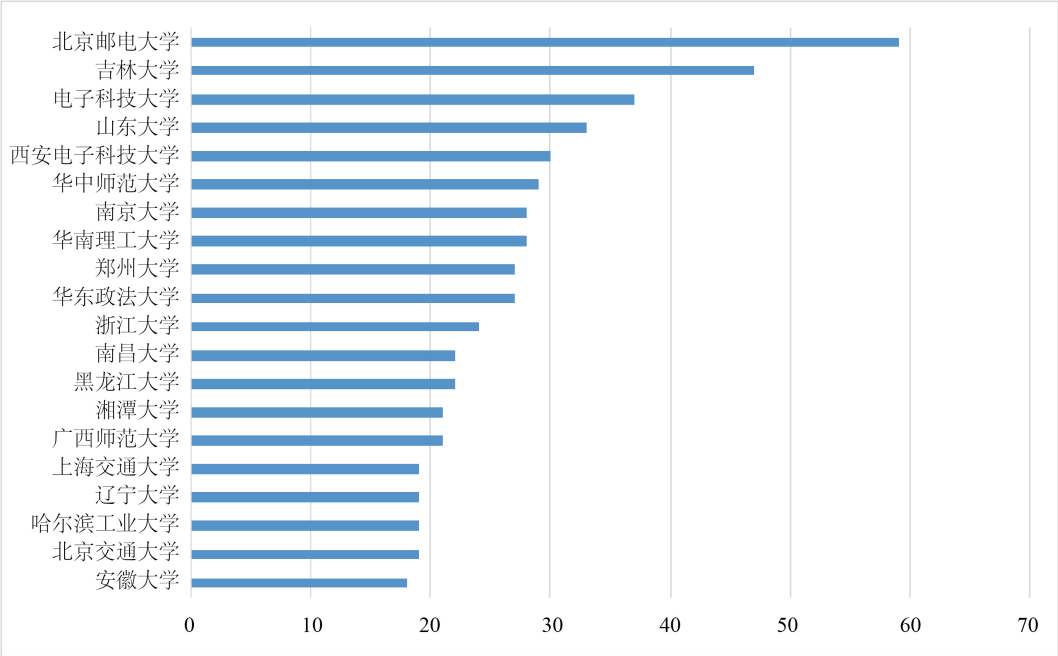

Figure 2. Ranking of degree awarding units of master's and doctoral degree thesis
图 2. 硕博学位论文学位授予单位排名

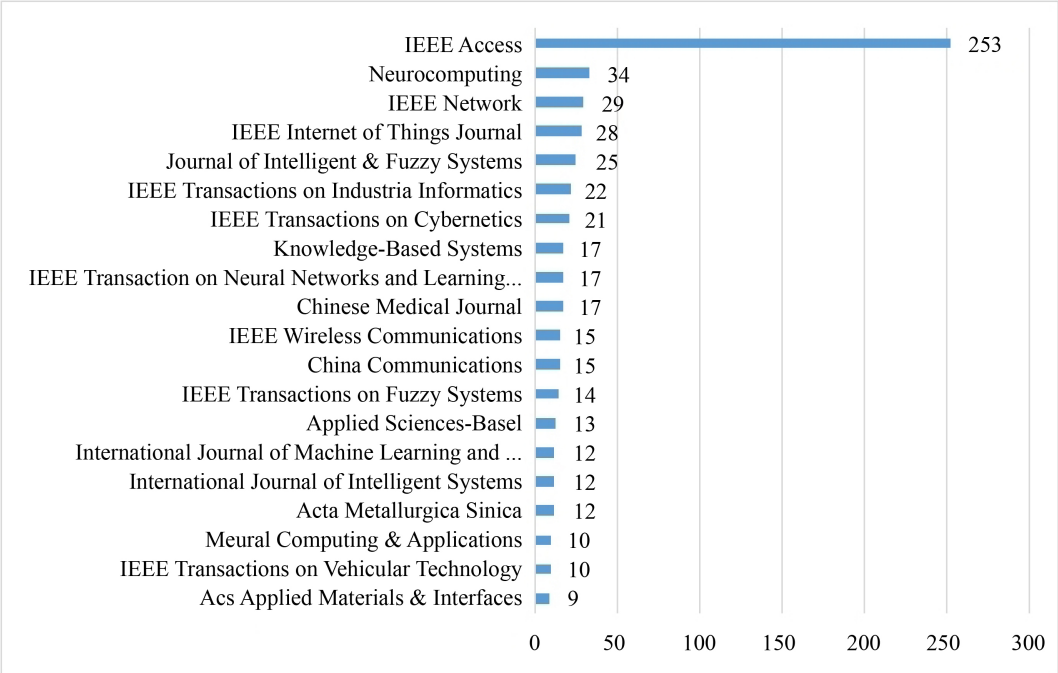

Figure 3. Ranking of foreign language journals containing Chinese authors' papers
图 3. 收录中国作者论文的外文期刊排名

Figure 4. Chinese journal paper author word cloud map
图 4. 中文期刊论文作者词云图

3.3. 关键词共词分析

利用软件 Python 对预处理后的关键词构造共词矩阵，根据共词矩阵解析文献关键词之间的网络关系数据，进一步通过 Gephi 软件计算关于网络分析的宏观指标和微观指标并进行可视化展示。

由于关键词出现的频率符合长尾分布，大量的关键词词频为 1，对词频较高的关键词进行共现分析才有意义，所以在共词分析中，需要自定义高频关键词[9]：1) 中文期刊文献：高频关键词限定最低频次为 5。2) 学位论文：高频关键词限定最低频次为 3。3) 外文期刊文献：高频关键词限定最低频次为 3。关键词共现网络宏观指标如表 1 所示。

Table 1. Macro indicators
表 1. 宏观指标

参数名称	中文期刊论文	中文学位论文	外文期刊
节点	86	88	65
边	303	253	533
平均度	7.047	5.75	16.4
网络密度	0.083	0.066	0.256
平均聚类系数	0.422	0.506	0.587
平均路径长度	2.398	3.298	1.733

根据表 1 可知：1) 中文学位论文整个网络的平均度数只达到了 5.75 左右，说明整个网络各个节点直接关联程度比较低，而外文期刊关联程度较高。2) 中文学位论文整个网络的密度为 0.006，相比来说是处于比较松散的状态，节点之间交流传递比较差。3) 外文期刊文献关键词的平均聚类系数为 0.587，平均路径长度为 1.733，相较其他两种更具有小世界性的特征。

除了表 1 中的指标参数，微观视角可以根据介数中心度、紧密中心度、离心率三个指标去度量每个节点上的关键词在网络中的作用[10]。本文以介数中心度为度量指标，指标越大越可以被认为是网络的中心节点或者说中心关键词。可以发现中文文献中指标前三位的关键词是一致的。外文文献中深度学习、机器学习关键词也在前三的位置。具体见表 2 所示。

Table 2. Micro perspective: Top 10 keywords of betweenness centrality index
表 2. 微观视角：介数中心度指标 Top10 关键词

中文期刊文献	中文学位论文	外文期刊文献
大数据	深度学习	Deep Learning
深度学习	大数据	Machine Learning
机器学习	机器学习	Task analysis
区块链	自动驾驶汽车	Sensors
算法	自然语言处理	Internet of things
5G	侵权责任	Training
图书馆	翻译策略	Computer Architecture
变革	法律规制	Neural Networks
数字经济	神经网络	Optimization
新型冠状病毒肺炎	算法	Resource Management

接下来，将基于微观指标，结合网络共现图对高水平文献的研究内容进行进一步探索。首先，需过滤一些意义不大的节点。由于很多关键词不是中心关键词，这些节点(关键词)与其他节点的关系并不紧密，即度很小。为了在可视化展示时更加清晰明了，本文设定节点的限制条件为：1) 中文期刊和学位论文文献：度最小为 4；2) 外文期刊文献：度最小为 10。其次，本文采用 ForceAtlas2 算法进行网络可视化布局，该算法具有运行速度快、处理的图形规模大的特点，在实际运行中，此算法形成的网络图更易进行社区发现。再次，将含有关键词信息的边文件和节点文件导入软件 Gephi，运行 ForceAtlas2 布局算法，采用 Louvain 算法计算模块度，判断是否适合进行社区发现，本文的文献模块度计算结果如表 3 所示。

Table 3. Modularity calculation results of literature
表 3. 文献模块度计算结果

	中文期刊文献	中文学位论文	外文期刊文献
模块度	0.529	0.708	0.445

在实际网络分析中，模块度若在 0.3~0.7 之间，说明聚类效果很好。由表 3 可知，三类文献有很好的聚类效果，均适合进行社区发现。最后，对无向网络共现图进行外观调整。以度为渲染方式，调整节点尺寸，根据社区划分选择区别度较大的颜色，以便更加直观的进行可视化展示。最终得到的三类文献的网络共现图分别如图 5~7 所示。

3.4. 基于摘要的 LDA 主题模型

本文对摘要进行 LDA 主题模型的搭建，以识别摘要中蕴含的主题，挖掘摘要中隐藏的信息，从而探索人工智能领域文献的热点研究方向。

首先，对文献摘要进行分词处理，将词组向量化，生成模型语料库。在这个过程中，由于有些词不具有重要意义，但在结果中占比较高，影响到了主题模型的权重判断，因此本文在常用停用词列表的基础上，增添了如“人工智能”、“技术”“研究”等词汇，将不想投入模型计算的一些高频词过滤掉，提高了模型提取的准确率。其次，使用 Jupyter Notebook 中 gensim 模块实现 LDA 模型的搭建，并基于变

分推断 EM 算法求解每一篇文档的主题分布和每一个主题中词的分布。再次，在 gensim 模块中使用层次狄利克雷过程(HDP)，自动确定适合语料库的主题个数。最后，选取最重要的 10 个主题，输出各个主题的前 30 个特征词，通过 pyLDAvis 进行可视化分析，动态显示各个主题下的特征词。可视化结果如图 8 所示。

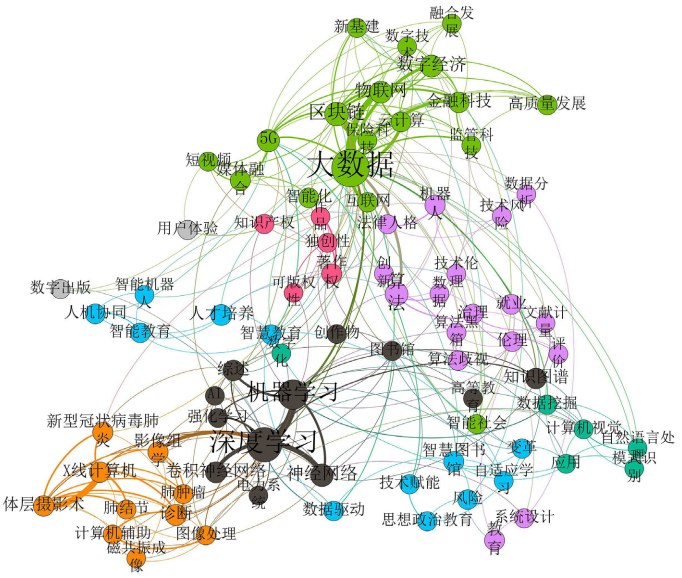

Figure 5. Chinese journal literature network co-occurrence atlas
图 5. 中文期刊文献网络共现图谱

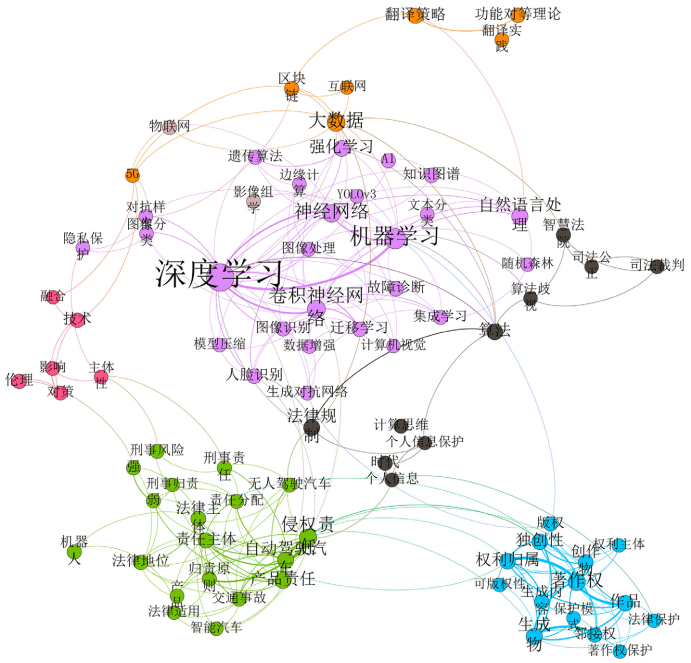

Figure 6. Network co-occurrence atlas of master's and doctoral dissertations
图 6. 硕博学位论文网络共现图谱

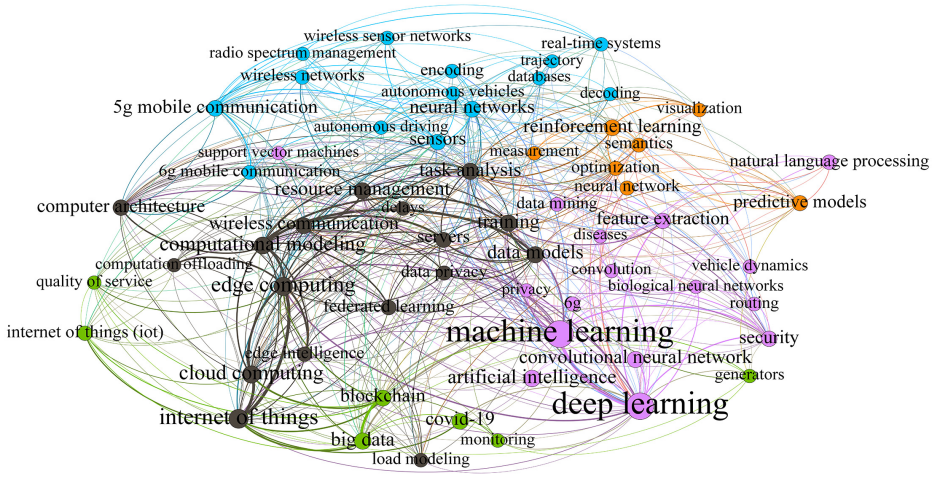

Figure 7. Co-occurrence map of Chinese authors and foreign periodicals
图 7. 中国作者外文期刊文献网络共现图谱

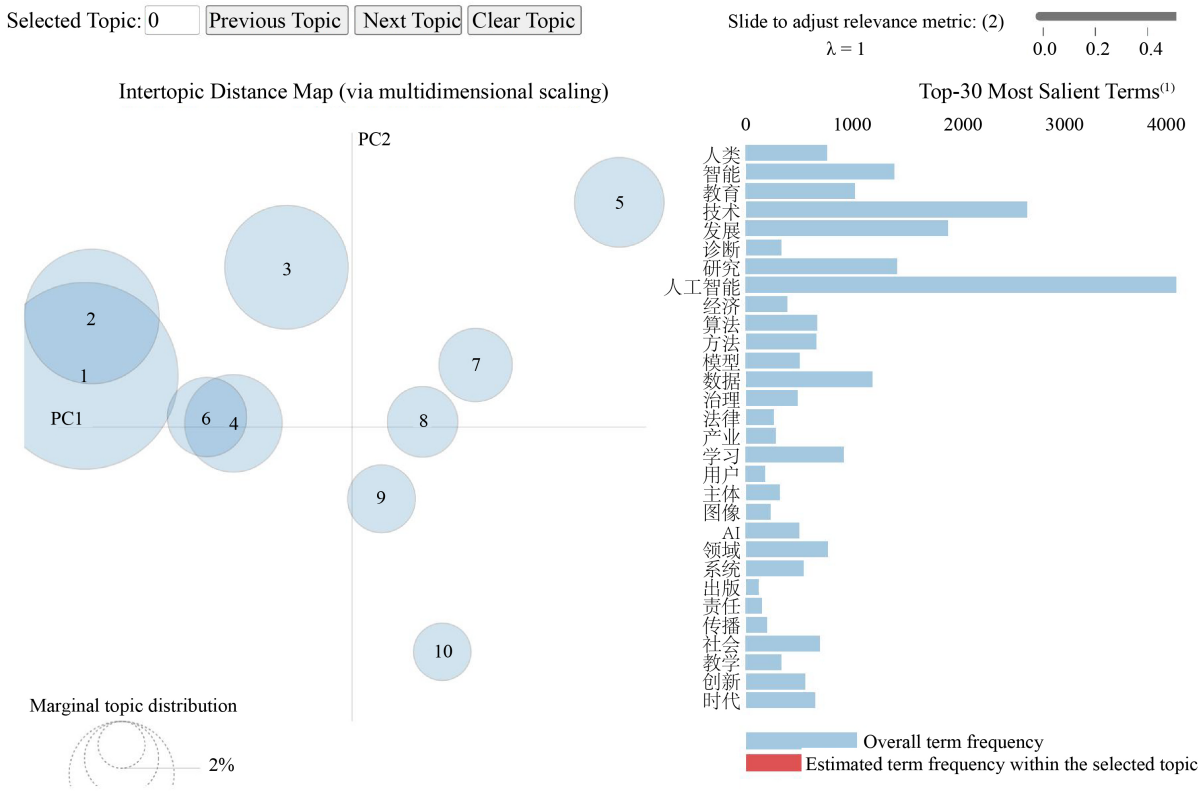

Figure 8. pyLDAvis visualization-theme dynamic diagram
图 8. pyLDAvis 可视化－主题动态图

本文将中文期刊论文、学位论文、外文期刊中文作者论文的摘要分词结果作为原始语料，调用 LDA 模型，得到了三种论文来源的各主题相关词汇。对三种来源文献的相同主题进行合并后，得到了近三年我国人工智能领域文献的研究热点，概括如下：

1) 技术融合促产业发展：利用大数据、人工智能、5G 通信、区块链、云计算等新技术，推动媒体

融合、金融科技、新基建、经济高质量发展。2) 人工智能技术应用的担忧与解决策略：包括隐私保护、算法歧视、教育公正、技术滥用、社会伦理、智慧司法决策风险、著作权、知识产权、作品独创性、可版权性等；3) 人工智能在教育领域的应用：主要聚焦于自适应学习、人机协同、智慧图书馆，加快推动人才培养。4) 人工智能在影像组学领域的应用：将体层摄影(CT)、X 线计算机、磁共振成像与人工智能结合，辅助肺结节及肿瘤的评估、新型冠状病毒肺炎的诊断、皮肤癌的分类、糖尿病视网膜病变检测、及骨骼影像处理等。5) 机器学习的技术融合与应用：热点方向为计算机视觉、自然语言处理及模式识别。

4. 结束语

本文运用文献的描述计量、文本挖掘等定性定量分析技术，针对我国人工智能领域的高水平文献，从文献的关键词和摘要两个方面共同分析人工智能领域的研究热点。通过对关键词进行共现分析，可以发现人工智能应用广泛，大数据相关不断火热，深度学习、机器学习算法在国内外理论研究中均占有主要地位；通过对摘要进行 LDA 主题模型的建立，挖掘近三年我国人工智能领域蕴含的研究主题，发现比较热门的主题包括技术融合促产业发展、人工智能技术应用的担忧与解决策略、人工智能在教育领域、影像组学的应用、机器学习技术的融合与应用五个主题，这对研究者快速发现人工智能领域的主题热点具有一定的参考价值。

参考文献

[1] 臧维, 张延法, 徐磊. 我国人工智能政策文本量化研究——政策现状与前沿趋势[J]. 科技进步与对策, 2021, 38(15): 125-134.

[2] 袁野, 于敏敏, 陶于祥, 等. 基于文本挖掘的我国人工智能产业政策量化研究[J]. 中国电子科学研究院学报, 2018, 13(6): 663-668.

[3] 李首骏.我国人工智能领域政策文本的量化研究[D]: [硕士学位论文]. 合肥: 安徽财经大学, 2020.

[4] 高长元, 张晓星, 张树臣. 多维邻近性对跨界联盟协同创新的影响研究——基于人工智能合作专利的数据分析[J]. 科学与科学技术管理, 2021, 42(5): 100-117.

[5] 李牧南, 王雯殊. 基于文本挖掘的人工智能科学主题演进研究[J]. 情报杂志, 2020, 39(6): 82-88.

[6] 吕一博, 韦明, 林歌歌. 基于专利计量的技术融合研究: 判定、现状与趋势——以物联网与人工智能领域为例[J]. 科学学与科学技术管理, 2019, 40(4): 16-31.

[7] 魏雪飞. 国内中小学人工智能教育研究热点及趋势——基于 CiteSpace 的文献计量分析[J]. 中国教育信息化, 2021(24): 6-12.

[8] 陶波, 陈敏. 中美医疗人工智能研究的比较分析[J]. 中国数字医学, 2018, 13(10): 35-38.

[9] 曲靖野, 陈震, 胡轶楠. 共词分析与 LDA 模型分析在文本主题挖掘中的比较研究[J]. 情报科学, 2018, 36(2): 18-23.

[10] 付鑫金, 方曙, 庞弘燊. 基于共词分析的我国情报学博硕士学位论文研究热点分析[J]. 情报科学, 2011, 29(11): 1722-1725.