DOI:10.13718/j.cnki.xsxb.2018.05.014

基于大数据分析的人工智能文献研究①

景 佳

西南大学 数学与统计学院,重庆 400715

摘要: 1992－2017 年间我国“人工智能”研究阶段性显著、发文量大、研究人员众多、研究阵地分散、各地区研究水平差异大、研究人员合作程度不够高,未形成较大的合作网络,综述性文献偏多、应用性文献偏少、人工智能研究成果的学科分布不均匀,本研究给出了相关建议和策略.

关 键 词: 人工智能;大数据;知识图谱;统计

中图分类号: G250.1 **文献标志码:** A **文章编号:** 1000－5471(2018)05－0080－06

随着信息技术的发展,人工智能已经成为世界各国关注的焦点. 2017 年“十九大”报告中提出推动互联网、大数据、人工智能和实体经济的深度融合. 同年 12 月,“人工智能”入选“2017 年度中国媒体十大流行语”[1],人工智能的迅速发展将深刻改变人类社会生活、改变世界. 因此,了解人工智能的发展态势,把握研究热点、难点问题显得十分必要.

1 数据来源

期刊最突出的特点是出版速度快,内容新颖,能迅速反映各领域的新信息,能为报道不断发展着的知识提供良好的条件[2]. 有鉴于此,笔者采用 CNKI 中国知网数据库为检索系统以“核心期刊＝Y”并且“主题＝人工智能”并且“关键词＝人工智能”为条件进行检索,检索时间范围为 1992 年 1 月到 2017 年 10 月,经过筛选、剔除、相似度计算后,保留有效数据 2397 条,在此基础上,对人工智能研究进行了探讨.

2 文献大数据统计分析与挖掘

笔者从文献发文时间、文献页数、文献作者(作者发文量、分布区域、合作度、合作率、高影响力文献作者分析)几个方面,运用统计分析和大数据挖掘对人工智能文献进行梳理.

2.1 研究阶段分析

一般来讲,某一学科或专业领域的研究通常会呈现出阶段性,人工智能的研究也不例外. 我们根据近 26 年人工智能论文数量来考量目前国内关于其研究的总体成果,通过不同时间段发表相关论文的数量将人工智能研究划分成若干阶段,针对不同阶段进行了深度分析.

我们先对所有 2397 条数据进行筛选,发现了 3 篇发表年份不详的文献数据,对整体数据进行了统计,具体结果见图 1.

由图 1 可知,仅从“人工智能”相关研究的文献数量来看,自 1992 年起,每年出版的文献平均数量起伏不定,其中,2004 年、2005 年、2007 年、2017 年均达到了 120 篇以上. 由此,我们将人工智能研究分为 4

① 收稿日期:2018－03－16
基金项目:中央高校基本业务费专项资金资助(XDJK2018C076).
作者简介:景 佳(1981－),女,四川巴中人,馆员,主要从事图书馆学研究.

个阶段：1992－1996 年，文献数量基本呈现阶梯型下降趋势；1997－2005 年，文献数量呈现阶梯型上升趋势；2006－2014 年，文献数量再次呈现阶梯型下降趋势；2015－2017 年，文献数量呈现直线上升趋势，从 66 篇跃升至 122 篇.

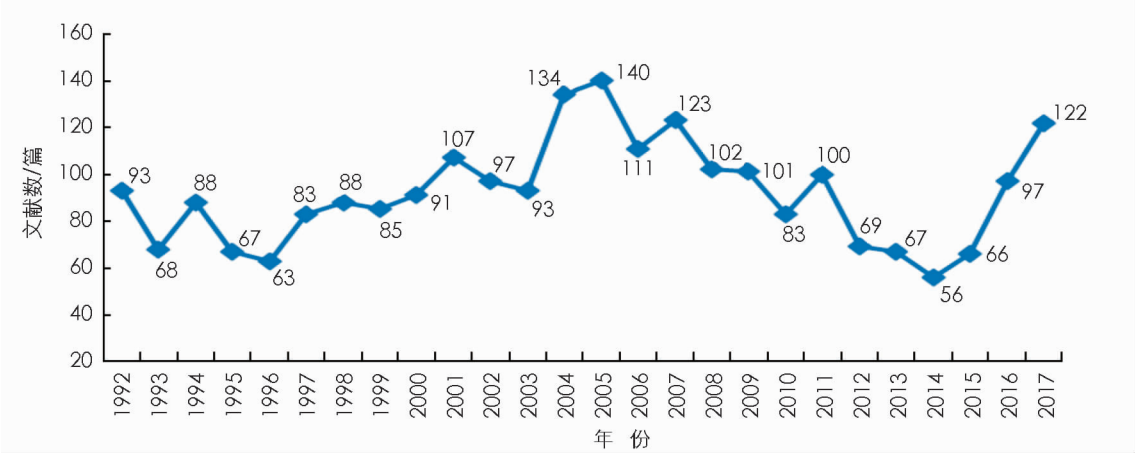

图 1　各年份“人工智能”文献数量统计

通过上述分析，“人工智能”已成为国内研究的热点问题，2005 年是人工智能文献出版最多的一年，之后研究热度虽有所减弱，但从 2015 年开始，热度明显增长，人工智能领域再一次获得了学者的重视.

2.2　文献页数统计

文献页数在一定程度上可反映文献对问题研究的深度，由于学科的差异性，有的文献侧重对原理的研究及实际应用，有的则侧重对理论、成果的介绍及概括，因此不同学科文献页数也存在差异. 对人工智能文献统计的结果见图 2.

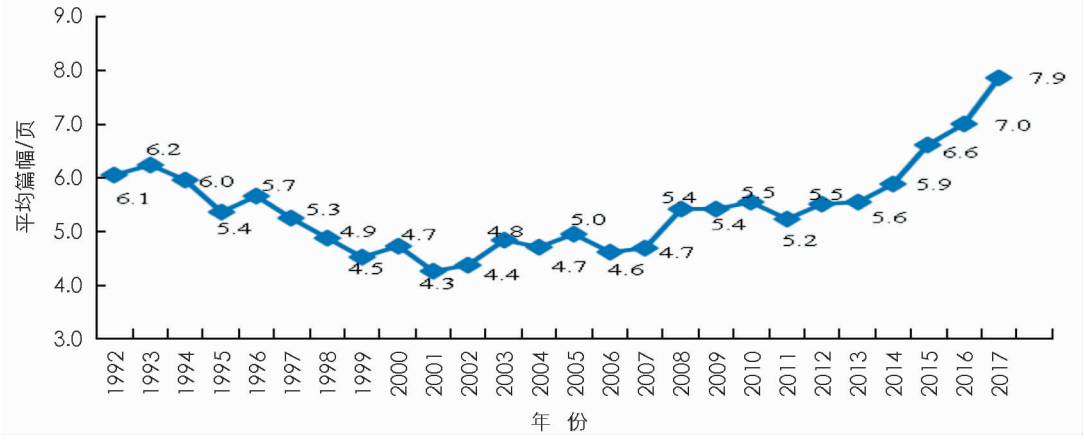

图 2　各年份文献平均篇幅

我们发现，2017 年文献平均页数最多(平均 7.9 页)，2001 年最少(平均 4.3 页). 1992 年至 2001 年，文献的平均页数总体呈下降趋势；2002 年至 2011 年，文献平均页数波动较大，但基本上还是有所上升；从 2012 年开始平均页数逐年明显上升.

对研究人工智能平均篇幅最长和最短的 10 门学科进行统计，结果见表 1.

从学科角度讲，表 1 反映了各个学科对人工智能本身的一种认可度. 从表 1 中我们发现，中国政治与国际政治、交通运输经济 2 个学科紧密联系人工智能，其平均篇幅均超过 10 页；而政治学、档案及博物馆、蚕蜂与野生动物保护 3 门学科相对较低，仅 2 页.

2.3　文献作者分析

科学计量学奠基人普莱斯曾经指出，人类进入了“大科学时代”. 合作论文往往是对科学劳动者合作与交叉的有力说明. 文献作者是科学研究工作的主要人员，文献作者的科研能力在一定程度上决定了该领域

科研成果的数量和质量. 因此，从学科发展来讲，需要积极发掘和培养新的科研人员，还应该对现有研究队伍的情况有全面、深入地了解和认识，对重点作者的研究动态进行实时关注[3].

表1 平均篇幅最长和最短的10门学科页码统计表

篇幅最长的10门学科	平均页数/页	篇幅最短的10门学科	平均页数/页
中国政治与国际政治	13.0	体育	3.3
交通运输经济	10.0	高等教育	3.0
保险	9.5	公安	3.0
诉讼法与司法制度	9.0	会计	3.0
中国文学	9.0	神经病学	3.0
海洋学	7.5	特种医学	3.0
基础医学	7.0	预防医学与卫生学	3.0
经济理论及经济思想史	7.0	蚕蜂与野生动物保护	2.0
肿瘤学	7.0	档案及博物馆	2.0
生物学	6.9	政治学	2.0

2.3.1 研究人员发文量统计

每一门学科出版论文的数量在一定程度上可以反映学科在所在领域的地位. 美国科学哲学家库恩认为，研究者的情况一定程度上决定了科学的发展[4]，研究者在某领域的发文量可以反映出其对该领域研究的贡献情况，对研究人员发文量的数据分析是有必要的.

表2给出了发文量位于前20的研究人员的情况. 从全部数据和表2统计可知，参与人工智能的研究人员数量高达5 051人次，其中发文量最多的高达27篇，有75人发文量在5篇及以上，但绝大部分研究人员发文量均在5篇以下，有4254人发文量仅有1篇. 这表明有很大一部分研究人员对人工智能领域产生过兴趣，但能够持续投入研究的人员数量较少，后续研究需要进一步加强.

表2 发文量位居前20位的研究人员发文量统计

作者姓名	发文量/篇	作者	发文量/篇
刘大有	27	李 健	10
欧阳丹彤	17	王生生	10
涂序彦	17	宗成庆	10
徐 波	15	刘 挺	9
欧阳继红	14	张 伟	9
张贤勇	14	周春光	9
姜云飞	12	朱永利	9
刘 群	12	蔡自兴	8
莫智文	11	蒋志华	8
潘云鹤	11	李占山	8

2.3.2 第一作者分布区域

研究人员的分布区域可以部分地反映该地区在相关领域的研究程度和水平. 我们对第一作者所在地区进行统计和分析(图3).

从图3可知，目前在人工智能领域进行研究的国内学者所在地区分布广泛，主要集中在北京、上海、长春、南京、武汉、西安、哈尔滨等地，这7个城市包含的文献数量均超过100篇，且北京的文献数量高达513篇，占全部文献的21.4%.

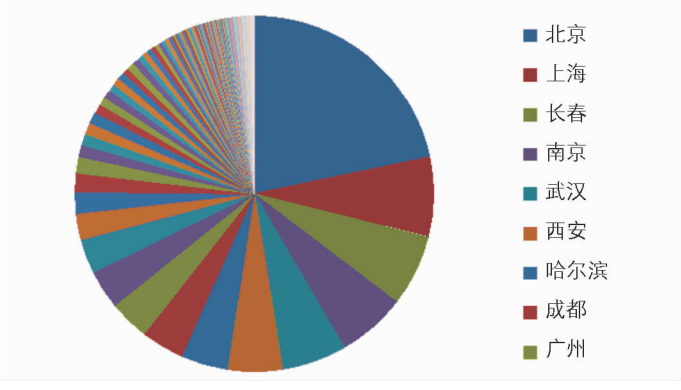

图3 第一作者所在地区分布

2.3.3 作者合作度及合作率

合作关系在科学研究中扮演的角色也日益重

要，且逐渐成为影响科学生产力的重要因素. 对合作关系的研究日益引起学者们的关注. 作者合作度及合作率是科学计量学和文献计量学中评估合作程度的常见指标，合作度是指作者总数与论文总数之比，合作率是指合作论文数与论文总数之比. 合作度与合作率反映了论文作者合作智能的发挥程度，数值越高，合作智能发挥就越充分.

在 2 394 篇文献中，共包含作者 5 051 人，合作论文数 1 924 篇，自 1992 年以来“人工智能”领域的研究人员之间总的合作度约为 2.11，合作率约为 80.37%. 具体每一年的详细合作信息见图 4.

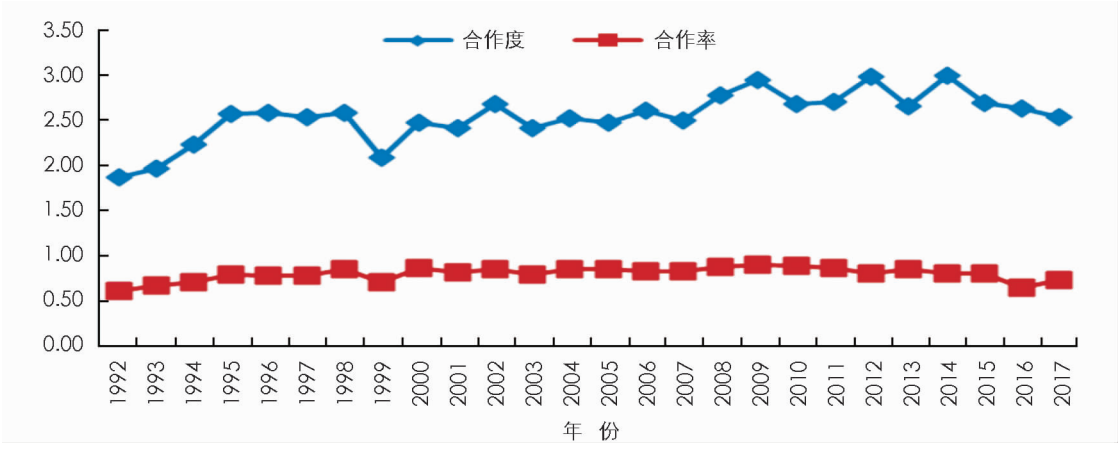

图 4　各年份作者合作度及合作率

从图 4 可知，我国人工智能领域的研究人员之间的合作度及合作率较高，相对比较稳定，表明我国对人工智能学科及相关交叉学科的重视. 而作者之间的合作率相对来说比较稳定，基本上保持在 80%左右.

2.3.4　高影响力文献作者分析

高影响力文献作者对学科发展具有非常重要的推动作用. 被引量、被下载量排名靠前的文献中，作者的合作度超过 2.5，合作率均高于 90%，同样显示出作者之间的相互合作更能充分发挥群体优势，不同合作者、不同学科之间相互取长补短，能显著提高科学研究成果的质量，从而增加文献的影响力. 在被引量、被下载量均排名前 50 的 27 篇文献中，作者人数达到 76 人，占全部文献作者人数(5 051 人)的 1.5%；排名前 100 的 42 篇文献中，包含的作者人数共计 124 人，占全部作者人数的 2.5%. 作为社会网络的一种，合著网络可以被理解为合著者及其合著关系的集合(图 5). 我们进一步通过绘制高影响力文献作者的知识图谱深度挖掘作者之间的合作情况(图 6).

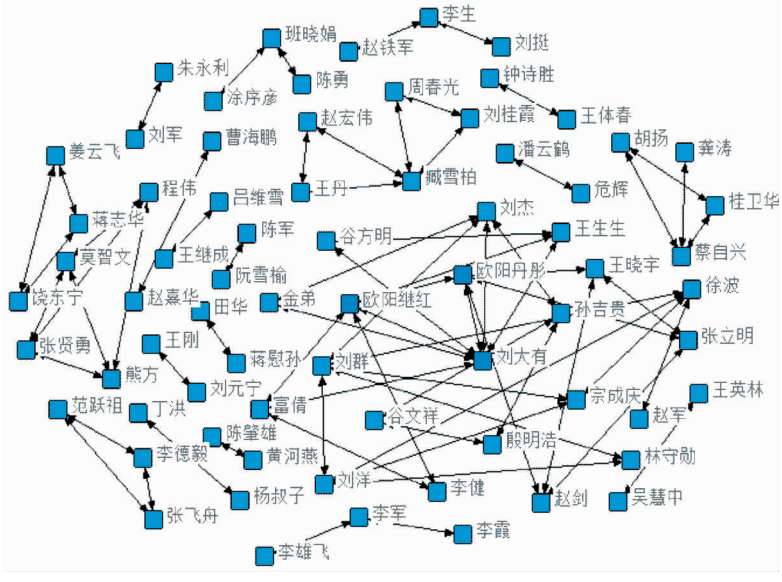

图 5　发文量前 100 位作者共现网络

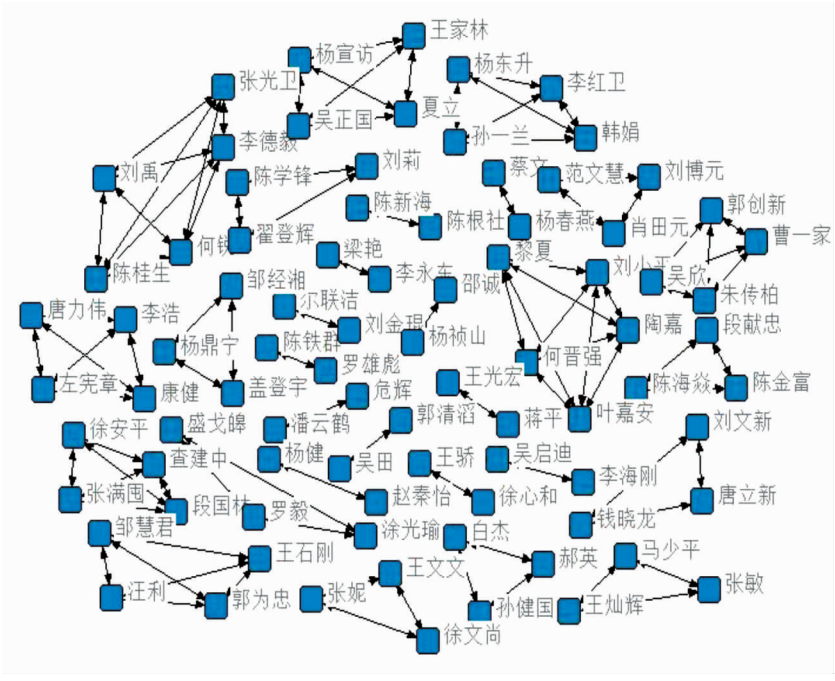

图6 被下载、引用量均位于前100的作者知识图谱

美国Drexel大学怀特(White)博士认为，作者共现频次越高，则作者学术相关性越强. 从核心作者共现网络(图5)和高影响力文献知识图谱(图6)来看，两个网络均较分散，没有形成核心的研究群体. 参与研究的作者数量众多，被下载量、引用量均位于前100的作者中有合作关系的有83位，有28个关系网络，有20个合作网络，其中2个网络包含作者5位，6个网络包含作者4位，9个网络包含作者3位，其余网络包含作者仅2位. 而共现网络中，1个网络包含21位作者，1个包含5位作者，2个包含4位作者，5个包含3位作者，11个包含2位作者，这表明高影响力文献的作者关系网络比较简单，并未形成实质性的核心研究团队，其研究成果影响力虽然较大，但作者之间的合作程度还有待进一步加强.

3 小结与建议

从近26年的人工智能文献的大数据分析与挖掘，我们清晰地看到，自1992年以来，我国人工智能研究存在的问题有：阶段性特点显著，发文量大，但不少研究浅尝辄止，对某个问题的研究不够深入；研究人员众多，研究阵地分散，各地区研究水平差异大；研究人员合作程度不够高、未形成较大的合作网络；综述性文献偏多，应用性文献偏少；人工智能研究成果的学科分布不均匀等. 这些问题需要研究机构、研究人员的加强合作、深化研究予以解决.

对此，人工智能研究人员应该充分认识到该领域未来对国家、社会发展的重要性. 我们需要通过加强地区、机构间的合作，形成中心化研究阵地，促进各个地区科研机构、高校之间的交流合作；促进研究人员之间的合作交流，互通有无，取长补短，形成核心研究群体；以产业发展为导向，重视应用技术研究，促进人工智能技术不断发展，构建良好的研究平台，不断优化我国人工智能领域的研究.

参考文献：

[1] 张 烁. 2017中国媒体十大流行语发布“十九大”“新时代”上榜 [EB/OL]. (2017－12－12)[2018－01－20]. www.xinhuanet.com/politics/2017－12/12/c_1122094339.htm.

[2] 叶 鹰，潘有能，潘 卫. 情报学基础教程 [M]. 北京：科学出版社，2006.

[3] 胡 臻. 基于普赖斯定律与综合指数法的核心作者和扩展核心作者分析 [J]. 西南民族大学学报，2016，42(3)：351－354.

[4] 库 恩. 科学革命的结构 [M]. 上海：上海科学技术出版社，1980.

[5] 邱均平，刘艳玲. 近 10 年我国合著现象的研究进展 [J]. 图书情报工作，2011，55(20)：11－14.

[6] 文庭孝. 专利信息计量学 [M]. 北京：科学出版社，2017.

[7] SAID Y H，WEGMAN E J，SHARABATI W K，et al. Social Networks of Author－Coauthor Relationships [J]. Computational Statistics & Data Analysis，2008，52(4)：2177－2184.

[8] WHITE H D. Pathfinder Networks and Author Co-Citation Analysis：A Remapping of Paradigmatic Information Scientists [J]. Journal of the American Society for Information Science and Technology，2003，54(5)：423－434.

On Artificial Intelligence Literature via Big Data Analysis

JING　JIA

School of Mathematics & Statistics，Southwest University，Chongqing 400715，China

Abstract：The CNKI China Knowledge Network database was used as a source to collect core journals' literature from 1992 to 2017，And we found some problems on Artificial Intelligence as follows in the past 26 years：the research of artificial intelligence in China has a significant phase，there are a large number of publications and relevant researchers，there are some scattered research sites，there are large differences on level of research among different regions，low level of cooperation，there are no major cooperation networks，there are many review literatures and few applied documents，and there are some uneven distribution of artificial intelligence research results，etc. At last，we gave some relevant suggestions and strategies.

Key words：artificial intelligence；big date；knowledge graph；statistics

责任编辑　张　枸