文章编号:1674-6864(2020)04-0032-06 DOI: 10.16508/j.cnki.11-5866/n.2020.04.006

特征工程研究领域发展趋势的可视化分析

马利星,胡 敏
(北京信息科技大学 信息管理学院,北京 100192)

摘 要: 为了解国内外关于特征工程的研究进展与态势、研究热点与现状,使用CiteSpace软件对中国知网CNKI和Web of Science核心合集数据库收录的研究主题为特征工程的173篇中文文献和555篇外文文献进行计量学分析与可视化处理。通过对年度发文量、研究国家、研究作者、文献关键词等进行分析,得出结论:自2015年以来本研究领域中外文发文数量持续增长。中国和美国的相关研究成果居多,约占文献总数量的64%。LeCun Y、Bengio Y、刘挺、林鸿飞是较有影响力的研究作者。国内的研究热点主要有协议识别、xgboost、深度学习。国外的研究热点主要有深度学习、迁移学习、实体识别。研究前沿是深度学习和因子分解机。

关 键 词: 机器学习;深度学习;特征工程;CiteSpace;可视化分析
中图分类号: TP 391.4 **文献标志码:** A

Visualization analysis of development trend of the feature engineering research

MA Lixing, HU Min
(School of Information Management, Beijing Information Science & Technology University, Beijing 100192, China)

Abstract: In order to understand the progress, trend, hotspot and current status of feature engineering research at home and abroad, the CiteSpace software is used to analyze and visualize 173 Chinese literatures and 555 foreign literatures of feature engineering collected in the Core Collection Database of CNKI (China National Knowledge Infrastructure) and Web of Science. Based on the analysis of the annual volume, research country, researcher, and keyword, it is concluded that the number of Chinese and foreign publications in this research field has continued to increase since 2015. China and the United States have the most relevant research papers, accounting for about 64% of the total literature. LeCun Y, Bengio Y, Liu Ting and Lin Hongfei are influential researchers in this research field. Domestic research focuses on protocol recognition, xgboost and deep learning. Foreign research mainly focuses on deep learning, transfer learning and entity recognition. The research frontiers are deep learning and factorization machine.

Keywords: machine learning; deep learning; feature engineering; CiteSpace; visualization analysis

0 引言

随着大数据时代的到来,人们对数据价值的挖掘越来越重视,相关研究工作如火如荼地进行着,数据挖掘算法在各行各业得到广泛应用。无论是人为提取特征的机器学习,还是自动学习特征的深度学习,特征都发挥着重要的作用。特征作为算法的输入直接影响着模型结果的好坏,能否获得有用特征至关重要。而特征工程的目的就是获取重要特征。特征工程保障了高质量的输入,有利于得到良好的

收稿日期:2020-04-14
基金项目:教育部人文社会科学研究基金(20YJC630056);北京市教委科研计划基金(SM201811232003);北京市社会科学基金项目(15JGC171)
第一作者简介:马利星,女,硕士研究生;通讯作者:胡 敏,女,博士,副教授。

输出结果。

特征工程是在实践中发展起来的关于构建有效特征集的一系列方法的综合,对于不同的研究问题,使用的技术略有不同。很少有人对此进行专门研究。特征工程常与算法一起使用,绝大多数研究是针对某个具体的研究问题,探究合适的特征工程方法,以便训练出更好的模型。查阅更多的特征工程相关文献,有助于提出创新性的特征处理方法。

为了系统全面地了解国内外关于特征工程的研究,本文采用文献计量学的方法,借助 CiteSpace 软件对相关中外文文献进行可视化分析,了解特征工程的研究进展情况,掌握研究的知识基础和热点问题,为特征工程的相关研究提供理论参考。

1　研究工具与方法

CiteSpace 是由美国德雷赛尔大学计算机与情报学学院陈超美教授开发的一款能够绘制科学知识图谱的信息可视化软件。自 CiteSpace 开发至今,其在科技论文、学位论文、学术专著等研究中得到了广泛应用,用户数量十分庞大。该软件有很多功能。作者、机构、国家的共现分析不仅可以得到各个节点的发文量,还能看出节点之间的合作关系。通过对词频、词语时间趋势、词汇的网络属性等关键词共现分析可以得到某领域的研究热点和趋势。参考文献或作者共被引分析可以反映某研究问题的知识基础和研究前沿[1]。本研究使用的软件版本为 CiteSpace 5.6.R2。本文利用该软件绘制了关于特征工程领域的研究国家、研究作者、文献关键词的知识图谱,并对图谱所反映出的特征工程研究态势与热点进行深入分析。

2　数据来源与处理

为了全面了解特征工程的研究情况,需要对国内外的特征工程研究成果进行分析。在收集相关文献资料时,选择从中国知识信息资源最丰富的数字化学习平台——中国知网 CNKI(China National Knowledge Infrastructure)中获取中文文献,选择从收录了全球最具学术影响力高质量期刊的数据库——Web of Science 核心合集中获取外文文献。在 CNKI 中,选择专业检索,输入检索条件 TI = '特征工程' OR KY = '特征工程',经过删除无关文献,得到 2006 – 2020 年的 173 篇中文文献,检索日期为 2020 年 3 月 4 日。在 Web of Science 核心合集数据库中,进行主题检索,输入检索条件"Feature Engineering",选择文献类型 article 和 review,经过筛选得到 2008 – 2020 年的 555 篇外文文献,检索日期为 2020 年 2 月 28 日。

为了满足软件使用要求,CNKI 导出的数据文本格式选择"Refworks",Web of Science 核心合集数据库导出的数据文本选择"其他文件格式"中的"纯文本"格式,文本命名格式为"download_XXX"。另外,CNKI 下载的数据需要进行格式转换。将处理好的数据加载到 CiteSpace 软件中,通过选择不同的功能绘制相应的知识图谱。

3　知识图谱绘制与结果分析

3.1　各年度中外文发文量对比分析

通过各年度发文量可以看出某研究主题的发展情况和研究热度变化。因此,本文对关于特征工程的 173 篇中文文献和 555 篇外文文献进行统计对比并绘制了柱形图,如图 1 所示。从图中可以看出,2006 – 2014 年间关于特征工程的研究成果非常少,2015 年开始发文量明显增加,且保持较高的增长率,外文发文量约为中文发文量的 3 倍,说明国外关于特征工程的研究较多。发文量增加的原因是因为基于机器学习的研究方法受到研究者们的青睐,在多个领域展开了使用机器学习方法解决学科领域问题的研究,特征工程作为构建复杂特征集的方法,应用于众多研究中。例如,文本情感倾向性分析的一类重要研究思路是采用有监督的机器学习方法,它的核心就是特征工程[2]。

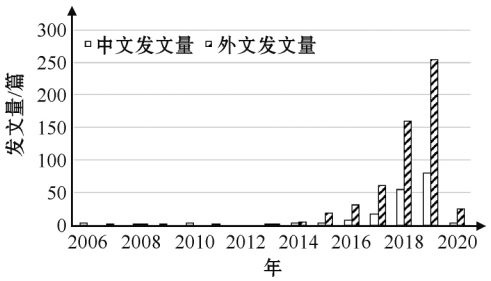

图 1　关于特征工程的各年度中外文发文量对比

3.2　主要研究国家分析

为了详细了解各个国家关于特征工程研究成果的多少及其重要程度,对获得的 555 篇外文文献进行国家共现分析,其结果如图 2 所示。节点大小代表发文量,节点文字的大小与发文量成正比,节点之间的连线粗细表示合作的强度。各个国家的发文量数据来自 CiteSpace 的统计结果。图中节点最大的国家是中国,发文量为 190 篇,占比约 34%,美国次

之，发文量为 167 篇，占比约 30%。发文量排名第三的国家是澳大利亚，仅发文 37 篇，与前两名差距较大。比较可得中国和美国对特征工程的研究较多，成果卓著。

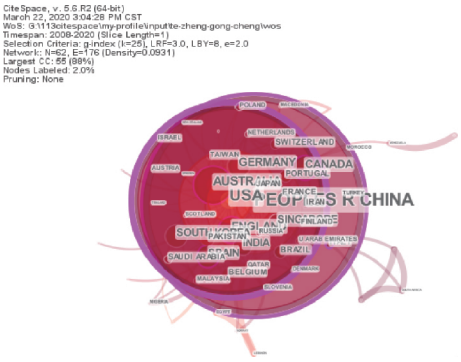

图 2 不同国家特征工程发文量共现分析

研究国家的影响力水平可以通过中介中心性这一重要的指标来衡量。根据 CiteSpace 提供的各个国家中介中心性的数据，对特征工程研究中介中心性排名前 10 的国家进行统计，如表 1 所示。美国的中介中心性排名第一，在特征工程研究领域占有重要地位。中国的发文量虽然最多，但中介中心性为美国的 7/10，说明我国需要提高发文的质量，加强创新性研究。中介中心性排名第三的国家是意大利，虽然发文量仅有 9 篇，但其研究成果的影响力超过了发文量较多的澳大利亚、西班牙、加拿大、英国。在表中，处于后三位的国家虽然中介中心性相同，但卡塔尔的发文量是法国和葡萄牙的一半，说明卡塔尔在该研究领域发表的文章影响力较强。由此可见，在进行科学研究时，评价研究作者或国家的科研水平，不能只关注发表论文的数量，更应该注重论文的质量，注重其研究成果的影响力。

表 1 关于特征工程研究中介中心性前 10 的国家

排名	国家	中介中心性	发文量
1	美国	0.53	166
2	中国	0.38	190
3	意大利	0.17	9
4	澳大利亚	0.15	36
5	西班牙	0.14	21
6	加拿大	0.13	25
7	英国	0.10	27
8	法国	0.09	12
9	葡萄牙	0.09	12
10	卡塔尔	0.09	6

3.3 作者共被引分析

通过研究论文被引情况，可以发现该领域较有影响力的研究作者。作者共被引分析是对论文和参考文献中作者之间的关系进行分析。为了解特征工程领域的研究作者论文被引情况，对获得的 555 篇外文文献进行作者共被引分析，结果如表 2、图 3 所示。表中共被引次数和中介中心性的数据均来自 CiteSpace 的统计结果。图中节点的大小代表被引用次数，节点文字表示作者信息。可以看出共被引次数排名第一的作者是 LeCun Y，排名第二的作者是 Bengio Y；从中介中心性的角度分析，Bengio Y 发表的论文在特征工程研究领域中影响力更大。这两位作者是深度学习领域最具影响力的科学家 Yann LeCun 和 Yoshua Bengio，他们和 Geoffrey Hinton 并称为深度学习三巨头，共同获得 2018 年图灵奖。在众多论文中，Yann LeCun 关于卷积神经网络、反向传播学习、文档识别的论文被引用次数较多，Yoshua Bengio 关于梯度消失、表示学习的论文被引用次数较多。另外，3 人在纪念人工智能提出 60 周年时，首次合作的综述文章《Deep Learning》被引次数非常高。该篇论文不仅阐述了深度学习的基本原理和核心优势，还详细介绍了 CNN、分布式特征表示、RNN 及其不同的应用，并对深度学习技术的未来发展进行展望[5]。

表 2 外文文献作者共被引次数前 10 名统计表

排名	作者	共被引次数	中介中心度
1	LeCun Y	123	0.19
2	Bengio Y	101	0.24
3	Krizhevsky A	87	0.13
4	Breiman L	82	0.17
5	Mikolov T	82	0.27
6	Hochreiter S	79	0.10
7	Srivastava N	70	0.04
8	Collobert R	65	0.17
9	Pedregosa F	53	0.03
10	Hinton GE	48	0.05

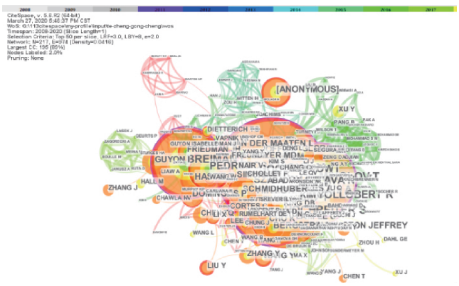

图 3 外文文献作者共被引分析图谱

为了解关于特征工程国内作者的研究情况，对 173 篇中文文献和 555 篇外文文献的国内作者发文情况进行统计汇总。借助 CNKI 的“分组浏览：作者”功能和 Web of Science 核心合集的“分析检索结

果: 作者”功能,统计了研究作者所在的机构和关于特征工程研究成果的被引次数,统计结果如表 3 所示。表中含有英文名的表示作者发表了关于特征工程的外文文献。表中研究成果被引次数最高的作者是刘挺,该作者是哈尔滨工业大学教授,入选国家“万人计划”科技创新领军人才,主要研究方向为自然语言处理和社会计算。在研究情感分析问题时,首次对中文微博语料进行细致的特征工程建设,提出了基于词典规则的情感评分新特征。曾与同单位的刘怀军和车万翔进行合作研究,使用特征工程解决中文语义角色标注问题,在英文语义角色标注特征的基础上,提出了一些更有效的新特征和组合特征[4]。另一位科研成果较多的作者是来自大连理工大学的林鸿飞教授。在关于特征工程的文献中,采用神经网络的方法进行生物医学文献的知识挖掘研究,具体研究问题有化学名与疾病间的关系、蛋白质间相互作用、生物医学事件触发识别等,为生物医学领域研究发展做出贡献。通过被引分析找到重要的研究作者可以迅速有效地了解某研究领域的研究进展与热点。

表 3　国内作者中外文文献被引次数前 10 名统计表

排名	作者	机构	被引次数
1	Liu Ting 刘挺	哈尔滨工业大学	185
2	刘怀军	哈尔滨工业大学	98
3	车万翔	哈尔滨工业大学	98
4	Lin Hongfei 林鸿飞	大连理工大学	50
5	Wang Jian 王健	大连理工大学	50
6	Yang Zhihao 杨志豪	大连理工大学	48
7	Zhang Yin 张音	军事医学科学院	37
8	Luo Ling 罗凌	大连理工大学	27
9	周国栋	苏州大学	19
10	白肇强	华南理工大学	13

3.4　关键词共现和聚类分析

文献的关键词反映了一篇文献的核心内容及重要信息,是对文献内容的高度概括和凝练。通过 CiteSpace 软件对文献中的关键词进行分析可以展现关键字共现网络图谱,并确定基于文献计量学视角的热点研究领域[5]。对关于特征工程的 173 篇中文文献进行关键词共现和聚类分析,结果如图 4 所示。聚类序号与聚类内包含的节点数量成反比,最大的聚类用“#0”标记,聚类名称用关键词命名。图中共有 13 个聚类,涵盖了特征工程研究的所属领域、模型算法与应用: ①聚类#0 数据挖掘、#2 机器学习、#4 深度学习主要是从宏观的角度表示特征工程所属的领域范围。数据挖掘使用机器学习算法和深度学习算法,无论是人为提取特征的机器学习,还是自动学习特征的深度学习,关于特征的研究都是其中必不可少的部分。②聚类#3 xgboost、#5 模型融合、#6 最大熵分类器、#7 决策树、#8 因子分解机、#10 信息抽取、#12 svm、#13 sequence to sequence 主要是从构建模型的角度说明特征工程经常与决策树、svm、xgboost 等算法一同使用。模型融合是采用两种或两种以上算法构建复杂模型解决某研究问题。最大熵分类器是自然语言处理领域进行语义角色标注常用的方法。因子分解机(FM,factorization machine) 模型因为能够有效解决高维数据特征组合的稀疏问题且具有较高的预测精度和计算效率,在广告点击率预测和推荐系统领域被广泛研究和应用[6]。sequence to sequence 简称 Seq2Seq,是一种在输入序列与目标序列长度不一致时采用的模型,可用于机器翻译、文本摘要、会话建模、图像描述等。③聚类#1 协议识别、#15 手势识别主要从研究问题的角度表明特征工程常用于解决网络协议和图像识别问题。其中,协议识别、xgboost、深度学习研究持续的时间最长,从 2006 年一直到 2019 年。因子分解机的研究从 2016 年持续至今,是特征工程研究领域的一个前沿问题。

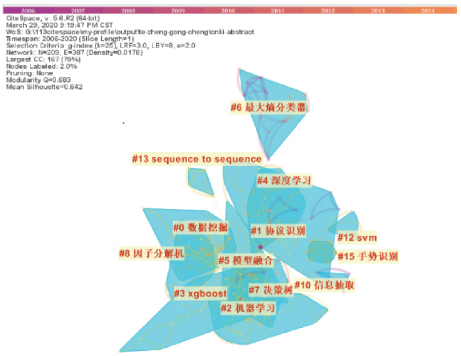

图 4　中文文献关键词聚类分析图谱

通过关键词聚类可得到多个研究主题。为了解各个研究主题的时间跨度,对 555 篇外文文献进行聚类分析,选择 CiteSpace 的时间线视图方式,结果如图 5 所示。图中展示了 2014—2020 年特征工程相关研究的发展情况,共得到 8 个聚类,聚类名称从文献的标题中提取。其中,聚类#4 中文网上健康咨询、#6 环境辅助生活系统、#7 元素组成可归纳为特征工程的应用场景这一研究主题。下面将详细地分析得到的 6 个研究主题。

1) 深度学习。深度学习是一类新兴的多层神经网络学习算法,通过组合低层特征形成更加抽象的高层表示(属性类别或特征) ,以发现数据的分布

图5 外文文献关键词聚类分析时间线视图

式特征表示[7]，解决了人工选取特征的繁复冗杂和高维数据的维度灾难问题。深度模型是实现特征学习的重要手段，深度学习和特征工程具有密不可分的关系，在深度学习的研究中都会涉及特征提取的问题。从图中可看出深度学习这一研究主题的时间跨度最长，从2014年开始持续至今。它是当前学术研究的热点之一，已经成功地应用于语音识别、图像识别等各个领域。

2) 迁移学习。从图中可看出迁移学习研究的时间跨度是2016－2019年。迁移学习作为一个新兴的研究领域，主要研究集中在算法方面。庄福振等[8]按照迁移学习方法采用的技术将其分为基于特征选择方法、基于特征映射方法和基于权重方法3大类。由此可看出迁移学习的大部分研究与特征工程相关。随着迁移学习研究的不断深入，特征工程理论将得到丰富和发展。

3) 实体识别。实体识别即命名实体识别，从图中可看出相关研究持续时间较长。命名实体识别是信息抽取、信息检索、机器翻译、问答系统等多种自然语言处理技术必不可少的组成部分。特征是影响命名实体识别的重要因素。多数情况下，命名实体识别系统使用的是基于统计和基于规则知识的混合方法[9]。其中，基于统计的方法与特征工程密切相关，对特征选取的要求较高。张祝玉等[10]在基于条件随机场的中文命名实体识别研究中，通过特征选取与组合的对比实验，得出在训练时应优先选择贡献度大的特征，同时还表明使用组合特征可以提升系统的性能。

4) 预处理机制。从图中可看出预处理机制研究的时间跨度是2016－2018年。虽然相关研究持续时间较短，但预处理是特征工程中的重要组成部分，包括针对单个特征的归一化、离散化、缺失值处理等，还有针对多个特征的降维和特征选择方法。在不同的研究问题中，预处理会影响到模型的结果，研究者们根据数据的特点采用合适的预处理方法，以便得到更好的结果。

5) 学习过程。学习过程包括深度学习过程和机器学习过程。两者都离不开特征工程。特征工程与深度学习的关系在前文已进行详细描述，此处不再赘述。对于机器学习而言，输入特征的质量直接影响着模型结果的好坏，通过特征工程可以得到有效的特征集，高质量的输入才能得到理想的结果。目前，机器学习已在各个领域得到广泛应用，但也面临着一些挑战。针对机器学习分类器存在的特征分类错误、精度低、过拟合等问题，Uddin Muhammad Fahim等[11]提出了一种新的算法融合和特征工程逻辑表单元的构造方案，使结果得到改善。

6) 特征工程的应用场景。特征工程作为提取有效特征集的方法论已广泛地应用于各行各业当中。近年来，关于网上健康咨询、元素组成、环境辅助生活系统国外的研究较多。在虚拟医疗行业，在线健康咨询已产生大量的医疗数据，一部分研究者通过特征工程和深度神经网络进行医疗实体识别研究，挖掘这些数据的价值，增强在线医疗的可持续性。在材料化学领域，一部分研究者将深度学习应用于元素组成的研究，通过深度神经网络模型自动捕捉不同元素之间的物理和化学相互作用和相似性，从而更准确、更快地预测材料的性能。环境辅助生活是一个新兴的多学科领域，针对子女无力看护老人这一社会问题，很多研究者利用人工智能技术设计了环境辅助生活系统，通过各种不同的监测仪器获取数据，对使用者的状态和环境对象进行分析并做出即时反应。其中，根据传感器收集的数据能否提取有效特征关乎着能否正确识别特定的活动，可见特征工程的研究至关重要。Ni Qin等[12]对环境辅助生活系统中需要识别的主要活动、传感器的设置、数据预处理和特征提取的方法进行了分类。

综上所述，特征工程与深度学习、机器学习密不可分，在协议识别、迁移学习、实体识别等研究中扮演着重要角色。特征工程作为提取有效特征集的方法论在各个领域得到广泛应用。

4 结束语

本文以特征工程的相关文献为研究对象，从中国知网CNKI获得173篇中文文献，从Web of Science核心合集数据库获得555篇外文文献，使用CiteSpace软件绘制多个知识图谱并进行分析解读。通过统计各年度中外文发文量发现自2015年以来特征工程研究领域的文献逐年递增，并保持较高的

增长率,这说明特征工程研究越来越受到研究者的关注。通过研究国家共现分析发现中国和美国的发文量约占总数的 64%,虽然中国的发文量高于美国,但研究成果的影响力较弱,这说明中国需要提高发文的质量。通过对研究作者们进行分析得出 LeCun Y、Bengio Y、刘挺、林鸿飞是研究领域内较有影响力的代表人物。通过对关键词共现和聚类分析得出特征工程在各个学科领域得到广泛应用,国内外所涉及的热门研究主题有所不同。国内的研究热点主要有协议识别、xgboost、深度学习。国外的研究热点主要有深度学习、迁移学习、实体识别。研究前沿是因子分解机和深度学习。

参考文献:

[1] 李杰,陈超美. CiteSpace: 科技文本挖掘及可视化 [M]. 北京: 首都经济贸易大学出版社, 2016: 77 - 78.

[2] 李泽魁,赵妍妍,秦兵,等. 中文微博情感倾向性分析特征工程 [J]. 山西大学学报(自然科学版),2014,37(04):570 - 578.

[3] LeCun Yann, Bengio Yoshua, Hinton Geoffrey. Deep learning [J]. Nature, 2015, 521(7553): 436 - 444.

[4] 刘怀军,车万翔,刘挺. 中文语义角色标注的特征工程 [J]. 中文信息学报,2007(01): 79 - 84.

[5] 全林发,陈炳旭,姚琼,等. 基于文献计量学和 CiteSpace 的荔枝蒂蛀虫研究态势分析 [J]. 果树学报,2018,35(12):1516 - 1529.

[6] 燕彩蓉,周灵杰,张青龙,等. 因子分解机模型的宽度和深度扩展研究 [J]. 软件学报, 2019,30(03):822 - 844.

[7] 陈珍,夏靖波,柏骏,等. 基于进化深度学习的特征提取算法 [J]. 计算机科学,2015,42(11):288 - 292.

[8] 庄福振,罗平,何清,等. 迁移学习研究进展 [J]. 软件学报,2015,26(01):26 - 39.

[9] 孙镇,王惠临. 命名实体识别研究进展综述 [J]. 现代图书情报技术,2010(06):42 - 47.

[10] 张祝玉,任飞亮,朱靖波. 基于条件随机场的中文命名实体识别特征比较研究 [C]. 第四届全国信息检索与内容安全学术会议论文集,2008: 118 - 124.

[11] Uddin M F, Rizvi S, Razaque A. Proposing logical table constructs for enhanced machine learning process [J]. IEEE Access, 2018, 6: 47751 - 47769.

[12] Qin Ni, Ana García Hernando, Iván de la Cruz. The Elderly's independent living in smart homes: a characterization of activities and sensing infrastructure survey to facilitate services development [J]. Sensors, 2015, 15(5): 11312 - 11362.