THUAITR

人工智能之学术搜索

Report of AI-powered Academic Search

2020 年第 2 期

清华大学人工智能研究院
北京智源人工智能研究院
清华-中国工程院知识智能联合研究中心

2020 年 5 月

目 录

图表目录

摘要

学术搜索（Academic Search）为科研工作者提供了一个可以从一个位置广泛搜索众多学科和资料来源学术文献的简便方法。随着人工智能（AI）技术不断引入，学术搜索产品的功能逐渐变得更强大、更智能，同时，结合 AI 技术的学术搜索产品也成为主要的发展趋势。

本报告以 AI 赋能的学术搜索为核心，在梳理学术搜索概念特征、发展历程、工作原理以及系统架构的基础上，重点研究分析了 AI 技术在学术搜索领域的具体应用情况、领域专家人才现状、典型产品的资源覆盖和功能特色，以及 AI 学术搜索领域的未来发展趋势，并探讨了学术搜索领域的市场主体如何才能更“智能”、更“聪明”、更“定制化”地为科研用户提供相关情报服务。

报告说明

一、重点/亮点

- 展示主流学术搜索产品中已引入的 AI 特色功能；
- 挖掘 AI 学术搜索领域专家学者并进行人才画像；
- 预测 AI 学术搜索技术趋势，为产品性能提升提出建议。

二、数据来源与研究方法

1.数据来源（详见附录 2）

（1）基于人工智能领域专家评议确定的领域最有影响力的会议和期刊；

（2）基于中科院期刊分区表中的计算机大类下的人工智能和信息系统两个小类的一区所有期刊会议论文。

2.研究方法

通过 AMiner 大数据平台对近 10 年（2009-2019 年）上述来源的论文数据进行挖掘，基于“学术搜索”相关的关键词库（详见附录 1），通过关键词智能匹配挖掘出所有相关论文。然后，基于这些论文，进行如下的进一步挖掘分析。

（1）通过文献分析，挖掘领域发展历程、技术特征；

（2）通过论文数据，挖掘领域关键词及其研究热度，进行技术趋势预测；

（3）通过论文作者相关信息，挖掘出该领域专家学者；通过抽取论文中所有学者信息，进行人才相关分析。

3.关键词抽取方法

（1）利用 AMiner 技术趋势产品搜索学术搜索相关的关键词，从中筛选出相关的关键词，再用所选出的关键词进行搜索，从中再次筛选出更多的相关的关键词；如此反复操作，扩展筛选并去重。

（2）通过相关论文的关键词进行拓展查找。最终得到学术搜索相关的关键词共计 112 个（详见附录 1）。

1

概述篇

科技信息资源是科技创新的物质基础! 当今时代, 数字化学术资源浩瀚丰富、多元互联。我国在《国家中长期科学和技术发展规划纲要》[1]、十二五科技发展规划[2]中都强调了科技情报大数据挖掘与智能服务的重要性。统计结果显示[3]，2009 年至 2019 年，中国科技人员共发表国际论文 260.64 万篇，排在世界第 2 位，数量比 2018 年统计时增加了 14.7%；论文共被引用 2845.23 万次，增加了 25.2%，也排在世界第 2 位。快速增长的科技文献规模已远远超出了个人的处理能力，亟需智能化的科技知识服务系统来辅助科研人员做分析挖掘。

针对此科研需求，国内外巨头纷纷推出学术大数据搜索和分析挖掘服务，尝试通过最新科技手段，助力科研工作者快速、准确、便捷地从互联网浩瀚的各类文献资源中查询出所需要的知识文献、掌握科技研究动态，并且能够从大量数据中发现隐含的、有价值的科技情报和科研规律，从而加快科技创新速度、提升创新研究效率。

1.1 相关概念

20 世纪 90 年代互联网应用初期，科研工作者们通过关键字搜索方式、利用如 Google、百度等这样的**通用搜索引擎**来进行信息或知识查找。这种搜索方式返回的结果虽然信息量大，但是存在查询结果不准确、采集深度不够、信息展示无序化等缺点。随着用户对某一特定领域、特定搜索信息需求的增加，垂直搜索引擎随之发展起来。

1.1.1 垂直搜索引擎

垂直搜索引擎（Vertical Search Engine ），又称为专业搜索引擎（Specialty Search Engines）、专题搜索引擎（Topical Search Engines），是搜索引擎的细分和延伸[4]。它针对某一特定领域、特定人群或特定需求来提供信息检索服务。

---

[1] 《国家中长期科学和技术发展规划纲要（2006—2020 年）》中华人民共和国国务院, http://www.gov.cn/jrzg/2006-02/09/content_183787.htm

[2] 科技部发布国家“十二五”科学和技术发展规划， 2011 年 07 月 13 日，http://www.gov.cn/gzdt/2011-07/13/content_1905915.htm

[3] 《2019 中国科技论文统计结果发布：从求数量到重质量 评价指标变化显著》，光明日报, http://www.gov.cn/shuju/2019-11/20/content_5453698.htm

[4] 许丽丽编著 网络信息资源检索与利用[M]．哈尔滨：黑龙江人民出版社，2008.12：59

垂直搜索引擎根据用户的特定搜索请求，通过对特定领域或行业的信息进行深度挖掘与分析整合、过滤筛选，以某种形式将结果返回给用户。其关键技术有聚焦、实时和可管理的网页采集技术，从非结构化内容到结构化数据的网页解析技术，精准全面的全文索引和联合检索技术，以及高度智能化的文本挖掘技术[5]。

垂直搜索引擎的应用方向很多，比如购物搜索、人才搜索、房产搜索、工作搜索、交友搜索等。基于文献检索的学术搜索就是一个细分的垂直搜索引擎应用。

1.1.2 什么是学术搜索

学术搜索是指专门为学者和科研人员服务，用于广泛搜索海量且涵盖各类学术期刊、会议论文、专利等学术文献的方法或平台。

简言之，学术搜索就是将互联网海量的、各种类型的学术资源进行收集整合后组成虚拟学术数据库，利用搜索形式为用户提供查询及搜索服务，使用户获得与搜索主题相关的论文、书籍、技术报告、专利等全球文献资源和学术科研信息。

学术数据库的元数据可以是图书馆的馆藏文献资源即数字图书馆中的信息检索系统，也可以是采购的商用数据库资源，在开放的互联网环境下还包括了与学术相关的文献、项目、专利、新闻等资源。在学术搜索系统中，这些元数据以学科、主题、人物、组织机构、基金等要素进行标引，构建出元数据仓储知识库，进而为用户提供各种学术文献资源的统一检索、资源揭示、资源调度与全文定位，让用户了解掌握到某领域最重要的学术文献或研究动态。

按照覆盖范围，学术搜索分为有综合性和专业性两类，前者面向各种学科类型的学术资源，后者则专门针对某类学科学术资源，例如用于搜索化学、生物医药信息的专业搜索。

1.1.3 学术搜索与网页搜索的区别

学术搜索和普通网页搜索的主要区别在于前者是垂直搜索而后者是通用搜索。通过**学术搜索平台，可以把普通搜索中大量无用的信息进行过滤**，更加有侧

[5] 刘俊熙，盛宇编著．计算机信息检索[M]．北京：中国铁道出版社，2009.08：134-135

重性、有针对性地找到相关的学术资源。比如在学术搜索中搜索某作者姓名，得到的搜索结果都是为用户呈现这个作者相关文献发表的期刊名称、期数、文献主要内容甚至包括支付方式等全面信息，而不会出现该作者的相关新闻网页。

学术搜索用户感兴趣的内容多是“深度”的学术科研信息。通过**学术搜索可以整合获得比普通网页搜索结果更多的有价值的内容。**网络资源中有很大一部分以深层/隐形网页的形式存在，由于存储格式或访问权限限制等原因而无法被普通的搜索引擎索引。这些深层网页多为数据库中的资源，其中很大一部分有同行评议等质量控制措施，具有较高的学术价值。传统的搜索引擎无法获取网络上这些以深层网页形式存在的资源。

1.1.4 学术搜索的特征与应用

学术搜索用于专门搜索学术资源，具有信息搜索结果查全率和查准率高、重复率低、相关性好、学术性强等特征和优势。

学术搜索主要应用于**查找与获取学术文献信息、科技查新、学术评价、作者合作关系、学者发现和推荐、会议与期刊查询和推荐**等多个场景。

根据我们对用户使用学术搜索产品的一项调研，发现国内用户最常用**的**学术搜索工具依次是谷歌学术、知网、百度学术，Web of Science（WOS）、微软必应学术、以及 AMiner 等工具。用户常用到的功能主要有以下几项。

- **找论文**：
  - ✓ 按标题（含子标题）、发表年份、关键词进行搜索。
  - ✓ 检索查找全文、获取论文、自动或批量下载文献原文。
- **找引用**：搜索论文引用情况、参考引用数、使用谷歌学术生成引用格式并导出、进行引文分析。
- **找专家**：搜索相关研究问题的学者及其信息，查找评审专家和领域相关专家最新工作，包括专家研究工作的影响力评价等。

- **找前沿/热点/新技术：**查找某一学科科研领域的问题答案、前沿科技或研究热点，以及在科研立项、新产品开发、技术引进、申请专利和鉴定科技成果等过程中，对科技内容进行新颖性分析或鉴证的技术查新。
- **找期刊/会议：**查找某一领域的学术期刊/会议资源特征以便进行投稿。

1.1.5 学术搜索引擎与学术数据库的区别

学术数据库是指在计算机可读介质上，使用一定方法将学术类信息组织起来的信息集合。其研究主要集中在检索方法、收录范围、检索结果分析比较等方面。

学术搜索引擎是指根据用户需求与一定算法，通过组织、管理和维护开放互联网中的学术信息，依托于如网络爬虫技术、检索排序技术、网页处理技术、大数据处理技术、自然语言处理技术等多种技术，用户经过一个检索入口便能快速获取网络学术信息。其研究主要探讨文献来源、检索功能、检索结果以及其与传统数据库的差别。

学术搜索是在开放互联网上海量科技相关信息而发展起来的，学术相关资源开放且丰富。在学术数据库已有的文献计量功能基础上，学术搜索引擎不断完善计量内容，开发了具有自身特色的计量产品。目前学术搜索引擎的文献计量功能也已发展到可视化和智能化阶段。

从数据来源上看，学术搜索引擎的引文数据来源多于学术数据库。这主要是由于学术搜索引擎是网络中学术文献信息的第三方集成平台，搜索结果仅提供网址链接，不提供文献全文，可以覆盖多种数据来源；而学术数据库通常是由数据库商通过与出版社等建立合作关系，将这些机构出版的期刊、图书等资源进行数字化处理，然后集成在数据库内部，数据来源较为固定和有限。

1.2 发展历程

学术搜索发展历程与计算机和网络技术发展密切相关。大致可以分为五个阶段（如图 1）：传统文献检索阶段（20 世纪 90 年代之前）、早期文献搜索阶段（20 世纪 90 年代-2000 年）、初期学术搜索阶段（20 世纪 90 年代末-2003 年）、开放互联网学术搜索阶段（2004-2009 年）、智能化学术搜索阶段（2010-至今）。

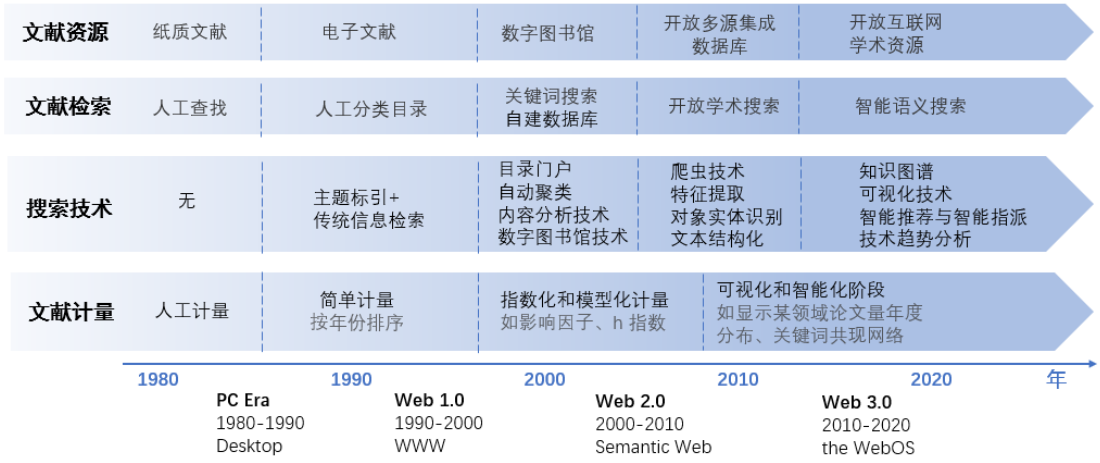


图 1 学术搜索发展历程

1.传统文献检索阶段（20 世纪 90 年代之前）

此阶段学术资源以纸质为主，搜索文献主要依靠人工编制目录作为索引来查找某一领域的论文图书等馆藏纸质文献资源。相关的文献目录主要按照学科主题、发表年份等指标进行人工简单排序，以方便用户查找所需文献。整个搜索过程没有引入计算机网络等相关技术。

2.早期文献搜索阶段（20 世纪 90 年代- 2000 年）

随着个人计算机和网络的应用，电子化数字化的学术文献逐渐兴起。20 世纪 80 年代出现了光盘数据库，推出了按文献发表年份排序等功能，以方便用户筛选出所需文献。此时的文献计量功能为简单计量阶段，主要表现为相关性排序、发表年份排序等；搜索技术主要采用的是**结合主题标引的传统信息检索技术**（Information retrieval）。

相比传统学术搜索阶段，这一阶段的用户也主要是查找论文，但可以按照主题、作者、期刊进行检索查找。

3.初期学术搜索阶段（20 世纪 90 年代末- 2003 年）

随着计算机和信息技术的发展，互联网应用已逐渐普及，数字化学术文献信息数量不断增多。一些机构逐渐创建了**自己的数据库、购买或共享多个学术资源库**。不同来源的数字资源可能具有不同的类型、格式或访问界面，被称为异构资源，存在于**各个相对独立的数据库**之中，允许授权用户或局域互联网用户访问。

这一阶段代表技术是**目录门户**（Directory Portals）、**自动聚类、内容分析以及数字图书馆技术。**学术搜索主要**以关键词搜索为主要特征**。

随着文献数据量的不断扩大，数据库的检索功能日益完善。网络版检索系统实现了简单文献分析计量功能，如索引词分析等。同时，随着学科的不断成熟，评价学术影响力的需求越来越大，文献的内容价值受到学者们的重视。不同数据源开始利用自身数据及引文数据等资源优势，构建了各种评价指标模型，如 Web of Science 中的期刊影响因子、中国知网的作者 h 指数、维普数据库的平均引文率、Scopus 的 CiteScore 等。此阶段的文献计量处于**指数化和模型化计量阶段**[6]**。**

学术搜索的应用范围已经不仅仅局限于查找论文文献，而是已经扩大到可以有限范围地找专家或评审、找期刊或会议、找研究前沿或热点等。

4.开放互联网学术搜索阶段（2004-2009 年）

互联网环境下的多源异构学术资源逐渐变得更加开放并且被集成。实际搜索中，用户也通常希望在同构的网络环境下“一站式”的访问并使用这些异构资源，而无需考虑这些资源的来源（网上免费资源、资源提供商的付费资源或图书馆的馆藏资源），甚至希望通过一次点击就可以获得资源的全文。

开放互联网的学术搜索主要采用了爬虫技术、特征抽取、对象实体识别、文本结构化等搜索技术，以学术资源为索引对象，涵盖互联网上的免费学术资源和以深层网页形式存在的学术资源，通过对这类资源的爬行、抓取、索引，以统一的接口向用户提供搜索服务。免费的学术搜索引擎如 Google Scholar 等在此阶段出现，为广大科研工作者搜索学术文献提供了方便。

在文献计量方面，学术搜索引擎在沿用已有的指数化和模型化计量基础上，不断完善计量内容，开发了具有自身特色的计量产品，并通过图形和图像等可视化方式呈现计量分析结果或特定的数据。例如，谷歌学术于 2012 年推出谷歌学术计量，用来评价各个领域杂志的影响力；百度学术自 2014 年成立以来推出了研究点分析、相关热搜词分析，具有深入计量文献的内容特征。

[6] 朱雯;陈荣;孙济庆;;多源数据的文献计量功能发展及其比较研究[J];图书馆理论与实践;2019 年 10 期

在这一阶段，学术搜索的应用范围继续扩大，用户可以在全球范围内查找论文文献、找专家或评审、找期刊或会议以及研究前沿热点等。

5.智能化学术搜索阶段（2010 年-至今）

随着计算机和信息技术的发展，尤其是 AI 技术兴起，免费的学术搜索引擎如 Google Scholar、Pub Med、百度学术等虽然为广大科研工作者搜索学术文献提供了方便，但是基于数据库技术构建的这些学术搜索引擎无法“理解”文献的内容。用户则希望搜索引擎能够帮助理解和处理这些数量庞大的数字化文献信息。

基于开放互联网学术资源，人工智能技术与学术搜索引擎的结合，使得“智能化”学术搜索和评价文献价值与学者学术影响力成为可能。这一阶段主要是以知识图谱、可视化技术、智能推荐与智能指派、技术趋势分析等搜索技术为代表。人工智能学术搜索，给科研人员的学术文献信息检索工作提供更多帮助，并大大提升工作效率。这使得学术搜索引擎不再仅仅限于为用户提供文章检索的简单功能，而是将深度学习技术用在信息筛选上，基于深度学习的检索系统能同时理解查询者的需求和文献的意思，以辅助科研学者更有效地检索学术信息。

随着知识经济时代的到来，不同数据源更加重视知识语义的挖掘，并且，在结果呈现方面更加多样化。如利用可视化技术显示某领域论文发展数量年度分布、关键词共现网络等，文献计量呈现出非常明显的**可视化和智能化特征**[7]。

在这一阶段，用户可以在全球范围内更加快速、更加精准地找到某一领域的论文文献、专家或评审、期刊或会议以及研究前沿热点等，更为重要的是，用户还可以通过学术搜索工具获得关于特定领域的文献、专家或期刊会议的智能推荐，甚至获知该领域技术趋势的预测，真正实现了智能化的学术搜索。

[7] 朱雯;陈荣;孙济庆;;多源数据的文献计量功能发展及其比较研究[J];图书馆理论与实践;2019 年 10 期

2

技术篇

AI 学术搜索技术既包含了传统的文献搜索技术，又涉及一些已引入应用的 AI 技术。其整体技术架构基本延续了传统的搜索架构。本部分在回顾传统搜索技术的基础上，重点梳理展示了相关的 AI 技术是如何提升学术搜索产品功能的。

2.1 学术搜索的工作原理

搜索引擎的基本工作原理（如图 2 所示）包括三个过程：首先在互联网中发现、搜集学术信息；同时对信息进行抽取和组织建立索引库；再由检索器根据用户输入的查询关键字，在索引库中快速检出文档，最后进行文档与查询的相关度评价，对将要输出的结果进行排序，并将查询结果返回给用户。

学术搜索引擎是应用于学术科研行业、专业的垂直搜索引擎。其工作原理与普通搜索引擎的基本类似，主要区别在于学术搜索引擎为用户提供的并不是上百甚至上千万相关网页，而是范围极为缩小、极具针对性的学术文献等具体信息。

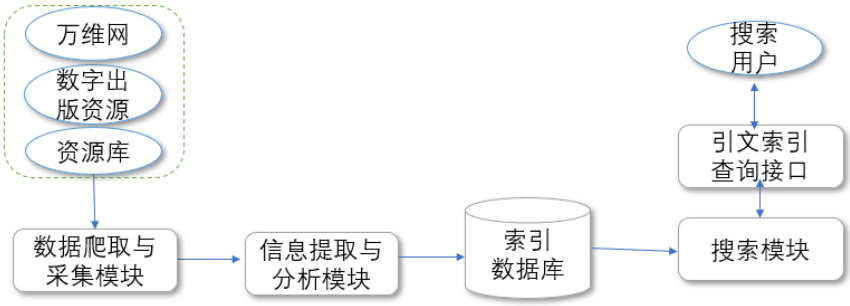


图 2 学术搜索引擎工作原理

学术搜索平台通常连接了作者、研究者、互联网学术资源（包括出版机构）、学术机构和用户等几大主体。主体之间是相互关联并可以相互转化的。例如，搜索用户可以访问出版机构资源以获取文献，也可以转化为作者通过搜索平台查找出版机构并投稿发表自己的科研成果。

2.2 学术搜索引擎系统架构

学术搜索系统架构基本相同。本报告以 2008 年的 AMiner 系统架构（如图 3 所示）为例进行说明。随着技术的不断迭代，如今的 AMiner 学术搜索系统已经引入了科技知识图谱、专家画像、认知图谱等新的功能技术模块，详情请其官网。

AMiner 通过对文献论文、专家学者社交网络、学术会议等科技大数据挖掘，利用知识图谱构建、隐含语义分析、情报快速匹配等技术，提供学术搜索服务。AMiner 学术搜索系统主要包括六大模块：**学术网络模型、社会网络存储、社会网络抽取、学者画像和专家发现、主题浏览、会议分析**。

通过**抽取研究人员资料**，系统可以自动为每个研究人员创建资料，包括基本信息，研究兴趣，社交圈和出版物记录。基于此元数据，系统支持在特定领域寻找和**发现专家**、权威期刊和会议以及有影响力的论文。此外，系统中的**社交网络搜索模块**支持用户查找研究人员之间的社交网络子图。**主题浏览**模块支持用户自动查找热门话题和话题的发展趋势。**会议分析**模块可以分析作者的国籍分布，被引用次数最多的论文和作者，以找出会议举办最成功的年份。

AMiner 学术搜索系统主要由采集、抽取、集成、存储和访问、建模和搜索服务六大技术模块组成。

1.**资源采集**：系统从互联网、论文数据库等多个资源中采集并获取学术资源。

2.**信息抽取**：自动从 Web 收集抽取研究人员的个人资料，以及从在线数字图书馆中抽取出版物信息；建立了专家人才档案进行**专家画像**，并且通过语义标注等技术，实现了**学术资源语义化**。

3.**语义集成**：以研究人员的姓名作为标识符，将所抽取的人员资料和出版物信息进行整合，通过概率框架处理名称歧义问题，集成后的数据存储在研究人员网络知识库中。

4.**存储和访问**：为 RNKB 中的抽取/集成数据提供存储和索引。具体来说，用 MySQL 进行存储，用反转文件索引的方法进行索引。

5.**建模**：利用生成概率模型同时对不同类型的信息进行建模，并为每种信息类型预估主题分布。

6.**搜索服务**：基于建模结果，提供专业知识搜索和关联搜索等多种搜索服务，以及作者兴趣查找、关于论文和引用的学术建议等其他服务。

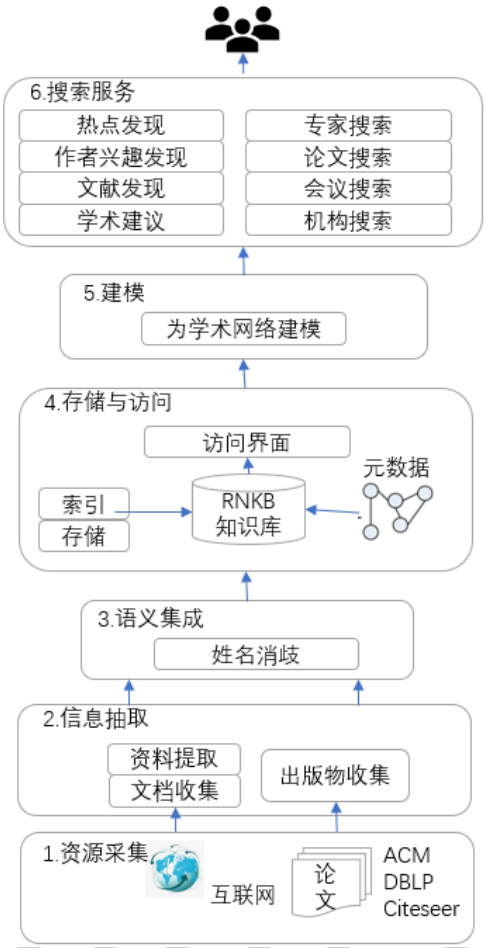


图 3 AMiner 专家与研究者学术网络搜索系统架构[8]

2.3 学术搜索主要技术

国内外各类科技资源网站，包括了论文、专利、人才计划、项目、重要学术会议和科技动态等类型。与普通文档资源网站的抓取技术一样，需要使用链接调度、资源抓取和解析提链等关键技术。不同之处在于，学术搜索还需要涉及到文献信息索引、作者识别、搜索结果排序以及推荐等相关技术。

近年来，人工智能、大数据、云计算等技术发展迅猛，已逐渐渗透到学术搜索领域。一些学术搜索系统中采用了当前最具前沿性的数据挖掘、自然语言处理、机器学习和知识图谱等人工智能核心技术，大大提升了学术搜索的产品性能和用户体验。

[8] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: extraction and mining of academic social networks. KDD.

2.3.1 信息抽取技术

搜索引擎通常使用网络爬虫技术自动从 Web 抽取研究人员个人资料，或者从在线数字图书馆中抽取出版物信息。将所抽取出信息之中的有效实体信息进行结构化存储，例如，论文标题、作者、出版物、出版年份等信息。

学术搜索系统的实体基本类似，例如，微软学术搜索（MAS）的核心是一个异构实体图（如图 4 所示），包括六种类型的组成实体：研究领域、作者、机构（作者所属）、论文、地点和事件。其中，论文和作者的实体发现数据来源主要包括两类：出版商的提要（例如 ACM 和 IEEE），以及由 Bing 编制索引的网页；研究领域（FOS）实体发现来自于内部知识库的数据；地点、事件和机构这三个会议相关的实体是从 Bing 索引的一些半结构化网站中收集的。

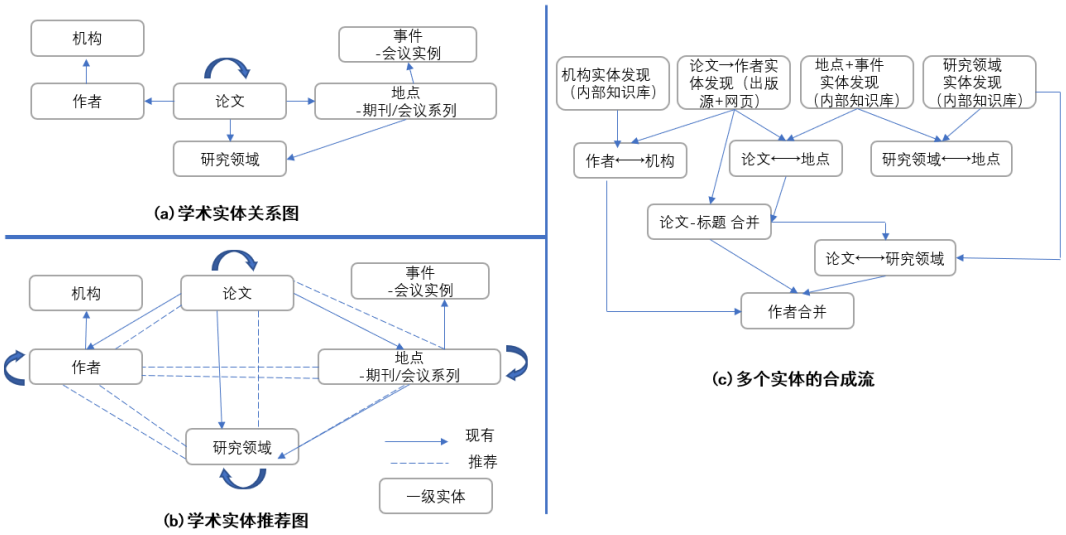


图 4 微软学术搜索的数据聚合和实体合并图[9]

正文抽取通常是基于文本密度和概率论两个因素进行的。文本密度越高的地方越有可能是正文，有效文本占总的文本比例越大的地方也越有可能是正文。关

[9] An Overview of Microsoft Academic Service (MAS) and Applications，Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-june (Paul) Hsu, Kuansan Wang，the International World Wide Web Conference Committee (IW3C2). 2015

于作者信息抽取，传统方法是从个人简历或网页中抽取个人信息、联系信息等[10]，如今是使用统一方法自动从网页上所标识的文档中抽取[11]。

2.3.1.1 基于机器学习和深度学习的信息抽取

机器学习（Machine Learning，简称 ML)是研究计算机如何模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能，最终使计算机具有人类智能。它是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多学科，也是人工智能的核心技术。

机器学习的处理系统和算法是主要通过找出数据里隐藏的模式进而做出预测的识别模式。

深度学习（Deep Learning，简称 DL）是机器学习领域中一个新的研究方向。深度学习是想通过模仿人脑的思考方式，建立类似于人脑的神经网络，来实现对数据的分析，按照人的思维做出相关解释，形成人们易于理解的图像、文字或者声音。深度学习是学习样本数据的内在规律和表示层次，这些学习过程中获得的信息对文字、图像和声音等数据的解释有很大的帮助。它的最终目标是让机器能够像人一样具有分析学习能力，能够识别文字、图像和声音等数据，解决复杂的模式识别难题。在 GPU 得到广泛应用后，深度学习基于人工神经网络算法，通过海量的数据训练神经网络，以达到机器学习的目的。

2.3.1.2 基于神经网络的信息抽取

神经网络技术可用于图片特征抽取。针对某个特定图片，通过卷积神经网络对图片进行特征抽取得到表征图片的特征，利用度量学习方法如欧式距离对图片特征进行计算距离，对图片距离进行排序，得到初级检索结果，再根据图片数据的上下文信息和流形结构对图像检索结果进行重排序，从而提高图像检索准确率，得到最终的检索结果。

[10] K. Yu, G. Guan, and M. Zhou. Resume information extraction with cascaded hybrid model. In Proc. of ACL'05, pages 499–506, 2005.

[11] Li, J., Tang, J., Zhang, J. et al. Arnetminer: expertise oriented search using social networks. Front. Comput. Sci. China 2, 94–105 (2008).

2.3.2 作者识别技术

作者识别主要进行三方面分析。一是**作者身份识别**(作者身份归属)，通过检查某一作者的其他作品来确定某一作者创作的作品的可能性。二是**作者身份描述**，根据作者的性别、教育、文化背景和写作风格，总结作者的特征，并生成作者简介。三是**相似度检**，对多篇文章进行比较，并确定它们是否出自同一作者之手，而不确定作者的身份，例如剽窃检测。

作者身份识别通常利用该作者的写作风格特征进行实验并计算出准确性。所采用的写作风格特征主要有**词汇特征、句法特征、结构特征**和**具体内容特征**。其中，词汇特征是基于单词和字符分析，句法特征表现为虚词、标点用法，结构特点是利用签名、个人文章组织风格进行识别，内容特异性特征分析一致使用的和与内容相关的关键词。

作者识别有两类计算方法（如图 5 所示）。一是**统计方法**，采用聚类分析和多维标度进行计算；二是采用**机器学习方法**，使用支持向量机。

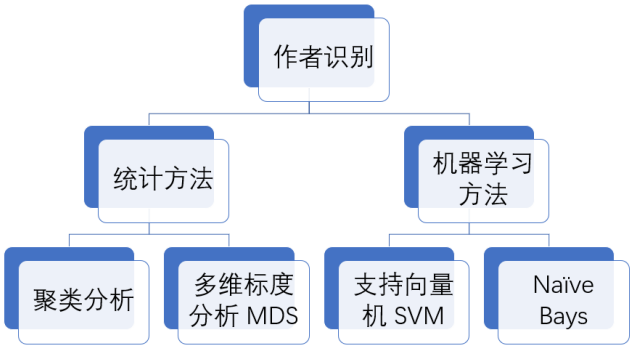


图 5 作者识别计算方法

2.3.3 命名消歧技术

不同数据源对同一真实实体的描述可能存在差异，因此需要研究基于来源置信度的多源冲突检测与消歧方法；每个数据来源对实体的描述也许比较片面，所以需要基于实体消歧结果实现信息补齐。

2.3.4 信息集成技术

检索系统根据建库时定义的可检索项提供多条检索途径，除了可检索如论文题目、作者、关键词等结构化数据外，还可检索全文、摘要、作者单位、期刊名、年份、期号等信息。各检索项和字段之间提供逻辑“与”、“或”、“非”关系的组配,整个查询信息的表达采用布尔表达、字段表达、自然语言表达相结合的方式。用户端的查询信息首先要进行分析处理，抽取出查询项索引、逻辑表达式或其他查询特征描述。

搜索平台需要进行**索引构建**。索引构建分为四个步骤：

（1）收集检索系统需要检索的所有文档，并构建出索引的文档集。

（2）把文本转化为词条（token），也就是把文本分割成一个一个的词。

（3）语言学预处理，通过词干还原（stemming）和词形归并（lemmatization）的方法，将词条进行归一还原，转为 normalized token，例如，将不同时态的单词转为其词根，将单复数名词统一转为单数形式。这些还原的 token 就是将要索引的词项（term）。

（4）对文档中的词项构建倒排索引。

2.3.4.1 基于自然语言处理技术的智能索引数据库

在文献信息处理阶段，采用自然语言处理技术**对各种文献源进行分析匹配，抽取关键信息，建立智能索引数据库**。匹配控制包括自由词匹配和概念匹配。自由词匹配是将用户提问与索引库中的索引项按照一定的检索模型进行匹配，将一系列包含该自由词页面的 URL 和摘要按查询相关度返回给用户。概念匹配又叫语义检索，是从词所表达的概念意义层次上来认识和处理用户的检索请求，匹配在语义上相同、相近、相包含的词语，旨在解决自然语言检索的同义和多义问题，且有助于提高查全率和查准率。

2. 3.4.2 基于人工神经网络的知识库管理

人工神经网络（Artificial Neural Networks，简称 ANNs）是一种应用类似于大脑神经突触联接的结构进行信息处理的数学模型。也简称为神经网络（NNs）或称作连接模型（Connection Model），在工程与学术界也常直接简称为“神经网络”或类神经网络。

人工神经网络是一种模仿动物神经网络行为特征，进行分布式并行信息处理的算法数学模型。这种网络依靠系统的复杂程度，通过调整内部大量节点之间相互连接的关系，从而进行信息处理。受人类大脑的生理结构启发的，人工神经网络具有离散的层、连接和数据传播的方向。

在搜索中，**神经网络应用于知识库管理**，就是搜索系统将逻辑推理与数值运算相结合，利用神经网络的学习功能、联想记忆功能、分布式并行信息处理功能，来解决检索系统的不确定性知识的表示与并行推理。

2.3.5 信息检索技术

早期搜索引擎使用的检索模型是**布尔模型。它**是基于集合论和布尔代数的一种简单检索模型，又被称为“完全匹配检索”（Exact-Match Retrieval）。它的特点是查找那些对于某个查询词返回为“真”的文档。

多数现代信息检索系统采用某种形式的**向量模型**的特征来扩展布尔模型，主要原因在于向量空间简单、快速，能产生更好的检索质量。另一种方法是用部分匹配和项权重的功能来扩展布尔模型。

1976 年，Robertson 和 Sparck Jones 提出了**概率模型**，利用了相关反馈信息逐步求精以期获得理想的查询结果。到目前为止比较常用的概率模型公式是 Robertson 提出的 BM25 公式。概率模型的主要缺点是对文本集的依赖性过强，且条件概率值很难估计。

1988 年 S.T. Dumais 等人提出了一种新的信息检索代数模型，即潜在语义分析（Latent Semantic Analysis）或者**潜在语义索引**（Latent Semantic Index）。它使用统计计算的方法对大量的文本集进行分析，从而抽取出词与词之间潜在的语

义结构，并用这种潜在的语义结构，来表示词和文本，达到消除词之间的相关性和简化文本向量实现降维的目的。

语言模型采用了一种逆向思维方式，该模型为每个文档建立了不同的语言模型，判断由文档生成查询的概率是多少，然后根据这个概率大小进行排序作为最终搜索结果。

随着文献资源的多语言性和用户所使用语言的多样性，**跨语言信息检索**（Cross-language Information Retrieval，简称 CLIR）变得越来越常见。跨语言信息检索是在对自然语言理解的基础之上，其关键问题是要使查询语言与文档语言在检索之前达成一致。使用户以一种语言提问，可以检索出另一种语言或多种语言描述的相关信息。在跨语言检索中主要涉及的关键技术有**计算机信息检索技术**、**机器翻译技术**和**歧义消解技术**。计算机信息检索技术完成提问与文档之间的匹配，机器翻译技术完成不同语言之间的语义对等，歧义消解技术则解决翻译过程中的多义和歧义问题。

2.3.5.1 基于知识图谱的检索模型构建

传统的学术搜索大多基于文档关键字匹配模型，为用户提供符合搜索条件的论文列表。以知识图谱构建智能搜索引擎，在传统的基于关键字搜索的基础上，提供了语义理解。而**借助于知识图谱，**学术搜索可以做到在用户键入查询词的同时理解用户的搜索意图，并根据查询意图做出特定的优化，从而**提高学术搜索的准确率和用户体验**[12]**。**

知识图谱是通过将应用数学、图形学、信息可视化技术、信息科学等学科的理论与方法与计量学引文分析、共现分析等方法结合，并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构，把复杂的知识领域通过数据挖掘、信息处理、知识计量和图形绘制而显示出来，揭示知识领域的动态发展规律。

学术文献中蕴含了丰富的信息，如研究主题、发表期刊、论文引用情况、作者、合作者关系网络、作者在某时段的工作单位或研究机构、受基金项目资助情

[12]汤庸，陈国华，贺超波, et al. 知识图谱及其在学术信息服务领域的应用[J]. 华南师范大学学报(自然科学版), 2018, 50(05):115-124.

况、刊物的覆盖内容和学术会议召开的时间及地点等。对这些信息进行深入分析，可构建出关系丰富的知识图谱。

2.3.5.2 基于计算机视觉的检索模型

计算机视觉可以看作是研究如何使人工系统从图像或多维数据中“感知”的科学。计算机视觉的基本原理是模仿人眼与人类视觉的立体感知过程，从两个视点观察同一景物，以获取不同视角下的感知图像，通过三角测量原理计算图像像素间的位置偏差，以获取景物的三维信息。计算机视觉系统通常可分为图像获取、摄像机标定、图像预处理、特征抽取、立体匹配和深度确定等六大部分。

将人工智能算法应用于视频图像的分类索引与检索中，系统将对视频图像的特征进行选取，包括了颜色直方图的计算、纹理的分析及应用运动跟踪，系统将会根据这些特征向量**对视频片断进行分类和检索。**

文献内的图片蕴含的丰富信息，在上下文语境、大小、命名等方面有着独特的特征，并具有信息需求、辅助滤检的功能，为用户快速了解文章内容提供了便捷。**在学术文献搜索中，以文献内图片为线索的检索从不同的思路出发，有利于检索者确定模糊的需求，获取更精确的检索结果。**因此，充分利用文献内的主题相关图像特征，对含主题相关图像的文献进行检索，并将图像缩略图在检索结果页面上予以显示，与搜索引擎自动生成的摘要、网页标题等共同辅助用户进行过滤检索和浏览检索，提高信息查找的效率。

2.3.6 排序技术

排序问题是信息检索和推荐系统等领域的核心问题之一。**传统搜索系统的结果排序**一般通过人工依据经验，去调整排序模型中所涉及到的一些参数，但这些经验参数不易调节且易产生过拟合。**方法大多依赖于人工经验**，由专家根据历史数据和待排序项的特征，通过组合一系列排序规则得到排序公式。

向量空间模型提出了一套可以进行部分匹配的框架。通过计算系统中存储的文档和用户查询之间的相似度，对所有检出的文档按照相似度进行降序排序。还有一种**模糊集模型**，它通过包含度定理计算其包含于文档的程度，根据这个包含度的大小来对检索出来的文档进行排序。

2.3.6.1 排序学习

排序学习 (Learning to rank) 是一个信息检索与机器学习相结合的研究领域。它利用机器学习方法在排序学习数据集上进行训练，自动产生排序模型，从而解决排序问题。和传统排序模型相比，排序学习的**优势在于对众多排序特征进行组合优化，对相应的大量参数自动进行学习**，最终得到一个高效精准、更加优化的排序模型。

排序学习方法根据其训练方式**分为三类**，逐点训练（pointwise），成对训练（pairwise）和列表训练（listwise）。其中，**逐点训练**的训练目标是优化对于一个文档的相关性分数估计，大部分的回归和分类机器学习方法都能用来训练逐点训练排序学习。**成对训练**排序学习每一次关注两个文档，给定两个文档，该排序学习会训练给出两个文档的相对顺序，一些比较流行的成对训练排序学习方法包括 RankNet, LambdaRank and LambdaMART。**列表训练**排序直接对整个列表进行训练，目标为直接优化列表的相关性排序，其训练目标可以是直接优化相关性排序指标，例如 NDCG 等，也可以是最小化刻画想要关注的列表的某一特性的损失函数，例如 ListNet 和 ListMLE 等模型。

2.3.6.2 深度学习

深度学习在基于内容的推荐中主要被用于从项目的内容信息中抽取项目的隐表示，以及从用户的画像信息以及历史行为数据中获取用户的隐表示，然后基于隐表示计算用户和项目的匹配度来产生推荐。深度学习算法模型与逻辑回归模型、支持向量机以及决策树类算法等传统机器学习算法模型相比，**主要区别体现在深度学习模型的网络结构包含更多更深的层级，并且明确强调特征表示学习的重要性。**该模型基于神经网络模型，却比简单的神经模型更为复杂，所处理的问题也更为复杂多样。最简单的深度学习模型莫过于多层感知机模型，其实深度指的就是隐层的数量，具有一个隐层的神经网络成为浅层神经网络，具有两层和两层以上的神经网络模型就可以称为深层神经网络模型也称为深度学习模型，将传统的一次非线性变换转换为多次的非线性运算组合构成了深度学习，深度神经网络模型比传统的神经网络模型具有更强的表示能力。

2.3.6.3 相关性排序算法

相关度排序是将查询结果按照与查询关键字的相关性进行排序，越相关的结果越靠前显示。

开源的全文检索引擎架构 Lucene 对查询关键字和索引文档的相关度进行打分，得分高的就排在前边。打分是在用户进行检索时实时根据搜索的关键字计算出来的，分以下两个步骤。

步骤一 计算词（Term）的权重

索引的最小单位是索引词典中的一个词（Term）。搜索是要从 Term 中搜索，再根据 Term 找到文档。

Term 对文档的重要性称为**权重**。影响 Term 权重有以下两个因素。

- **Term Frequency** (tf)：指此 Term 在此文档中出现了多少次。tf 越大，说明此词(Term)在文档中出现的次数越多，也说明此词(Term)对该文档越重要。
- **Document Frequency** (df)：指有多少文档包含此词（Term）。df 越大，说明此词越普通，不足以区分这些文档，也说明此词的重要性越低。

步骤二 根据词的权重值，计算文档相关度得分

此外，在搜索时，通过设置 boost 值（默认为 1.0f），可以对文档中的某个**域**（field）进行加权。这样，在进行组合域查询时，如果匹配到加权值高的域，则最后计算出的相关度得分就高，在搜索时匹配到的相应文档就可能排在前边。

2.3.6.4 基于自然语言处理技术的排序

自然语言处理 NLP 技术、语义分析技术，将用户的信息需求与文本进行概念匹配，返回给用户所感兴趣的信息，使检索结果更全面和准确，并根据一定的算法计算相关度并进行排序，把与用户查询最为相关的结果排列在前面。**自然语言处理能够提高查询结果排序的质量。**

2.3.6.5 基于机器学习与深度学习的论文影响力评价

学术搜索引擎 Semantic Scholar 具有评价论文影响力的 AI 功能。基于**深度学习技术**，该搜索引擎通过从论文中挑选出最重要的关键词和短语可以判断文章

所论述的具有评价主题，亦可以从论文中抽取图表，将它们呈现在检索结果中，帮助用户快速理解论文内容；也可以辨别一篇文章引用的参考文献是否具有重要的参考价值，基于此评价论文的学术影响力，可以快速获得重要文献。

2.3.7 推荐技术

2.3.7.1 基于关联规则的推荐

基于关联规则的推荐（Association Rule-based Recommendation）是以关联规则理论为基础。关联规则是用于在海量数据中挖掘出其背后隐藏的事务项之间的联系，通过数据之间的联系中的有用内容来获取更大的利益。假如两个项目事务之间存在一定的关联，那么其中一个项目事务可以以一定的概率作为前提条件来推断另一个项目事务。**基于关联规则的推荐技术重点在于其关注用户行为之间的关联模式**。

2.3.7.2 基于内容的推荐

基于内容的推荐根据内容信息和用户的偏好之间的相关性来向用户推荐信息。该技术一般通过类别或特征标签选择来获取用户的需求和喜好，然后通过信息过滤来获取更有价值的信息。

用户的兴趣模型常取决于所用的学习方法，比较常见的有决策树、神经网络、贝叶斯分类器、聚类等。基于内容的推荐技术的关键之处在于对项目的理解程度。目前大部分基于内容的推荐技术通常是对文本信息进行研究。

2.3.7.3 协同过滤算法

协同过滤是指通过群体的行为来找到某种相似性(用户之间的相似性或者标的物之间的相似性)，通过该相似性来为用户做决策和推荐。协同过滤技术被广泛用于预测用户兴趣偏好的应用领域。

协同过滤主要包括两种推荐算法：一是**基于物品(标的物)的协同过滤，**就是计算出每个标的物最相似的标的物列表，为用户推荐用户喜欢的标的物相似的标的物；二是**基于用户的协同过滤，**就是将与该用户相似的用户喜欢过的标的物的标的物推荐给该用户（而该用户未曾操作过）。

将用户对标的物的评分（或者隐式反馈[13]）构建成一个**用户行为矩阵**，矩阵的某个元素代表某个用户对某个标的物的评分（如果是隐式反馈，值为 1），如果某个用户对某个标的物未产生行为，值为 0。其中，行向量代表某用户对所有标的物的评分向量，列向量代表所有用户对某个标的物的评分向量。**行向量之间的相似度就是用户之间的相似度，列向量之间的相似度就是标的物之间的相似度。**

相似度的计算可以采用 cosine 余弦相似度算法来计算两个向量（例如，行向量或者列向量）之间的相似度：

$$sim(v_1, v_2) = \frac{v_1 * v_2}{\|v_1\| \times \|v_2\|}$$

深度学习目前被广泛应用于协同过滤推荐问题中。基于深度学习的协同过滤方法主要是将用户的评分向量或项目的被评分向量作为输入，利用深度学习模型学习用户或项目的隐表示，然后利用逐点损失（point-wise loss）和成对损失（pair-wise loss）等类型的损失函数构建目标优化函数对深度学习模型的参数进行优化，最后利用学习到的隐表示进行项目推荐。

2.3.7.4 组合推荐

实际应用中，为了达到更好的推荐效果，通常将多个推荐方法混合使用，也就是使用组合推荐（Hybrid Recommendation）。目前比较流行的组合方法是**特征组合**（Feature combination），它组合来自不同推荐方法的数据信息，一种推荐算法产生的数据信息被另一种推荐算法所采用。例如将协同过滤的产生信息作为增加的特征向量，然后在数据集上采用基于内容的推荐算法。

2.3.7.5 基于知识的推荐

基于知识的推荐（Knowledge-based Recommendation，简称 KB）是一种特定类型的推荐系统。它借助于领域本体、表达语义知识，增加了项目之间的关联信息。基于知识推荐系统的交互性很强，系统需要主动询问用户的需求，然后返回推荐结果。基于知识的推荐系统不需要用户评分数据就能推荐，推荐结果是以

[13] 隐式反馈，是指浏览、点击、播放、收藏、评论、点赞、转发等，以及任何不是用户直接评分的操作行为。

用户需求与产品之间的相似度或者明确的推荐规则而进行。知识获取则需要知识整理工程师将领域专家的知识整理成为规范的、可用的表达形式。

2.3.7.6 可解释性推荐

可解释性推荐是解决原因问题的个性化推荐算法，它不仅为用户提供建议，还提供解释，使用户或系统设计人员了解推荐此类项目的原因。通过这种方式，它有助于提高推荐系统的有效性、效率、说服力和用户满意度。近年来，可解释的推荐方法已被采用。

2.3.7.7 基于数据挖掘技术的智能推荐

数据挖掘（Data Mining）技术是一个跨学科的计算机科学分支。它是用人工智能、机器学习、统计学和数据库的交叉方法在大规模数据中发现隐含模式的计算过程。

在学术搜索领域，数据挖掘技术主要应用于搜索内容的智能推荐。某些 AI 学术搜索系统具有基于用户偏好和兴趣进行内容推荐的功能，即学术资源推荐系统。该系统中设立了一个描述用户偏好数据的**用户信息需求库**。通常，给每个用户建立一个用户描述文件，刻画用户的特征和彼此间的关系，通过用户描述文件能够准确地了解用户的兴趣，跟踪用户的兴趣和行为。

要形成用户描述文件，**需要通过数据挖掘技术获取数据，从中总结出用户的个性化特征**。通常而言，用户兴趣的获取途径一是用户自己制定提供；二是系统提供示例让用户选择，并通过对用户选择的实例文档进行分析来确定，当然也可以从用户或者系统与用户的交互过程来判断用户的兴趣；三是系统自动采集，并通过收集用户的浏览行为和服务器的日志进行数据挖掘，分析出用户的浏览行为、习惯，从而判断用户的兴趣，从而更好地提供个性化推荐服务。

此外，**数据挖掘技术还用于学术搜索系统用户和用户访问路径识别[14]。**根据搜索网站的拓扑结构图，系统对访问的用户进行识别判断。如果用户请求的某个页面不能从已访问的任何页面到达，则判断这是一个新用户；通过用户是连续请求页面，则识别其对服务器的一次有效访问，并可以获得用户在网站中的访问行

[14] 钟克吟. 基于混合推荐的学术资源推荐系统的服务模式与数据挖掘[J]. 图书馆学研究, 2013(11):58-61.

为和浏览兴趣。通过路径填充，获得用户完整的访问路径，然后经过片断识别而正确地识别用户有意义的访问路径。

2.3.7.8 基于机器学习和深度学习技术的智能推荐筛选

传统的过滤主要有基于包的过滤、基于应用的过滤和基于文本的过滤等几种。引入人工智能技术能够识别文档内容、进行不同学科论文的文本自动分类，从而实现智能化的过滤，可以让研究人员更方便地获取学科某方向的信息或学科的发展方向与趋势。

将深度学习技术应用在信息筛选上，**基于深度学习的检索系统能识别**用户意图、**理解用户的需求和文献的意思**，自动地进行个性化的搜索、筛选、信息推送，为科研工作者省去更多筛选的工作[15]。

基于深度学习的个性化推荐技术核心在于特征的非线性交互表达学习，即常用特征中不同纬度组合下的表达与学习，其中就包括三个特征层的组合：异质特征交互、同质特征交互、内容特征交互。

异质特征交互是根据场景数据中不同域的特征，比如年龄、性别、访问 IP 等异质特征，进行推荐预测，又包含特征域结构未知和特征域结构已知两种不同的问题，典型场景 CTR 推荐、社交推荐等。

同质特征交互是根据用户的历史行为，抽取并学习特征，并基于此进行相应的预测，典型应用场景如时序推荐。

内容特征交互是对所推荐物品的内容进行建模，一般延用并改进常用的图像、文本处理方式，典型应用场景社交媒体的照片推荐、新闻推荐。

当前深度学习在推荐系统研究中的应用可以分为**五个方向**：一是**基于内容的推荐**。利用用户的显式反馈或隐式反馈数据、用户画像和项目内容数据，以及各种类型的用户生成内容，采用深度学习方法来学习用户与项目相似的项目推荐给用户。二是**应用在协同过滤中**。利用用户的显式反馈或隐式反馈数据，采用深度学习方法学习用户或项目的隐向量，从而基于隐向量预测用户对项目的评分或偏好。三是**混合推荐**。利用用户的显式反馈或隐式反馈数据、用户画像和项目内容

[15] 谢智敏，郭倩玲. Semantic Scholar: Academic Search Engine Based on Deep Learning 基于深度学习的学术搜索引擎——Semantic Scholar[J]. 情报杂志, 2017, 036(008):175-182.

数据，以及各种类型的用户生成内容产生推荐，模型层面主要是基于内容的推荐方法与协同过滤方法的组合。四是**基于社会网络的推荐**。利用用户的显式反馈或隐式反馈数据、用户的社会化关系等各类数据，采用深度学习模型重点建模用户之间的社会关系影响，更好地发现用户对项目的偏好。五是**情景感知推荐**。利用用户的显式反馈或隐式反馈数据，以及用户的情境信息等各类数据，采用深度学习模型对用户情境进行建模，发现用户在特定情境下的偏好。

2.3.8 基于自然语言处理 NLP 和语义分析的用户交互技术

自然语言处理（Natural Language Processing，简称 NLP）是人工智能和语言学领域的分支学科。此领域探讨如何处理及运用自然语言，特别是如何编程计算机以成功处理大量的自然语言数据。

NLP 的基本任务包括正则表达式、分词、词法分析、语音识别、文本分类、信息检索、问答系统（如对一些问题进行回答或与用户进行交互）、机器翻译等。常用的模型则有马科夫模型、朴素贝叶斯、循环神经网络等。

一般而言，通过自然语言处理，探讨如何处理及运用自然语言；然后，通过自然语言认知，让电脑“懂”人类的语言；以及通过自然语言生成系统，把计算机数据转化为自然语言；通过自然语言理解系统，把自然语言转化为计算机程序更易于处理的形式。

在学术文献搜索阶段，自然语言检索接口**允许用户以自然语言的方式和机器交互**，接受用户自然语句输入的查询，让系统分析用户的自然语言提问，并通过人机交互推断出其真正需求。

3

人才篇

通过挖掘人工智能领域和信息搜索领域的国内外顶级期刊会议论文的作者大数据，本部分主要呈现了那些研究探索如何利用 AI 技术赋能学术搜索以不断提升搜索功能和用户体验的国内外专家学者现状。

3.1 AI 学术搜索领域的学者总体现状

通过 AMiner 大数据平台对 2009 至 2019 十年期间的来源期刊会议（详见附录 2）的所有论文数据，利用“学术搜索”相关的关键词库（详见附录 1）进行匹配挖掘，再通过抽取论文中所有学者信息，挖掘出该领域的专家学者，进行人才相关分析。

结果显示，全球学术搜索领域学者数量共计 7,262 位，从年度发展趋势上看，总体呈现上升趋势，且近 5 年来上升速度较明显，如图 6 所示。中国学术搜索学者数量总体保持稳定，约占全球领域学者总量的 10%。

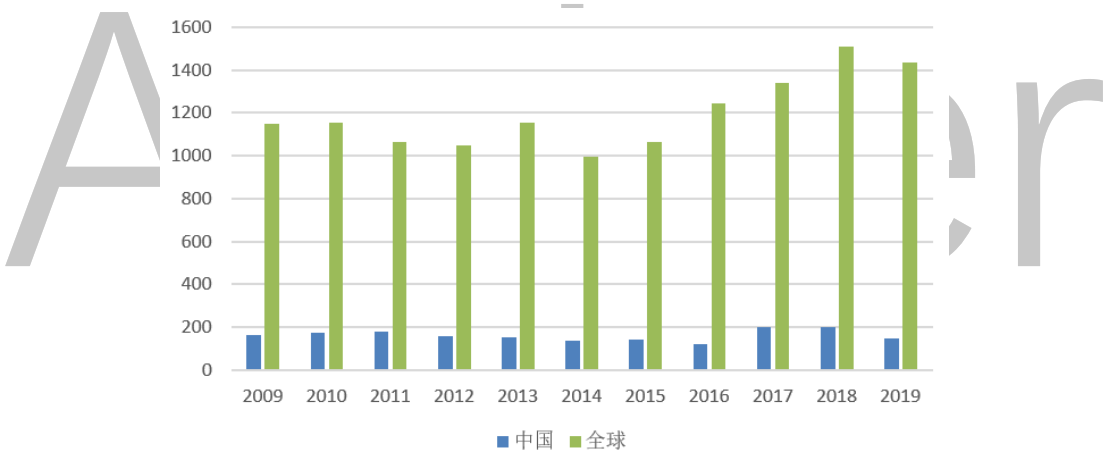


图 6 2009 至 2019 年领域学者数量趋势

3.1.1 学术水平情况

结果显示，全球领域学者的 h-index 均值 13.2，其中 h-index 低于 10 的学者量占比最多，为 58%；h-index 大于 20 的学者量占比 23%。这些学者人均论文发表量 55.3 篇，每篇论文平均被引频次 44.3 次。可能由于学术搜索是一个应用较强的领域，领域学者在学术论文发表方面并不十分活跃。

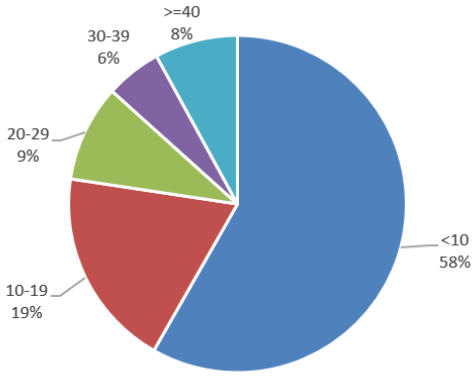


图 7 领域学者 h-index 值分布

3.1.2 学者国家分布

领域学者覆盖全球 54 个国家，遍布亚洲、欧洲、大洋洲和美洲 4 大洲（如图 8 所示）。从地域角度看，近一半的领域学者集中在北美洲，主要分布在美国地区；亚洲的领域人才也较多，主要分布于中国、新加坡、印度及日韩等地区；欧洲中西部也有一定的学者分布；其他诸如南美洲、非洲等地区的学者数量稀少。

图 8 领域学者国家分布

注：根据学者当前就职机构地理位置，绘制出领域学者的地域分布图。其中，不同图标颜色代表不同地区的学者，图标大小代表学者数量。

在中国地区，领域学者主要分布在北京、上海和深圳，安徽、陕西、湖北等省份也有少量分布，如图 9 所示。

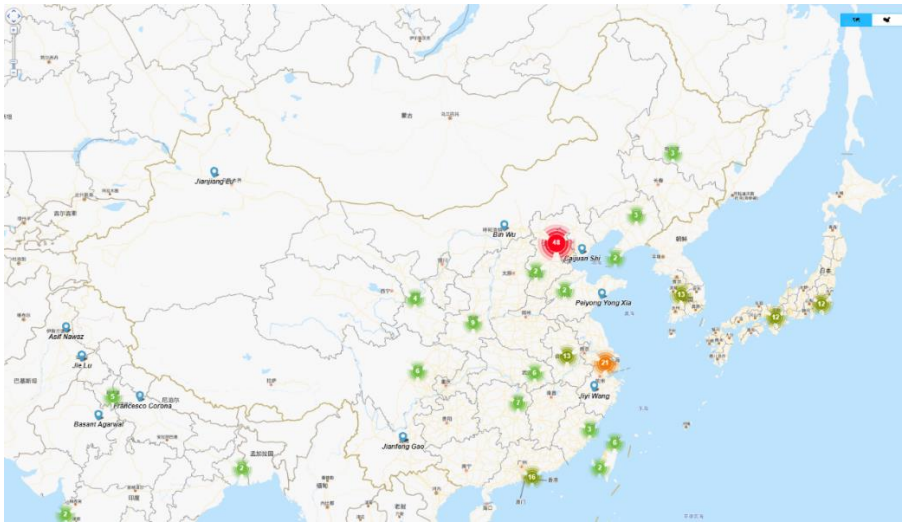


图 9 领域中国学者城市分布

注：根据学者当前就职机构地理位置，绘制出该领域学者的地域分布图。其中，不同图标颜色代表不同地区的学者，图标大小代表学者数量。

3.1.3 学者数量前十的国家

学术搜索领域学者数量前十的国家之中（如图 10 所示），美国、中国的领域学者数量分别位居第一、第二，均超过千人，其余国家领域学者数量均在千人以下。英国位居第三，领域学者数量为 312 人。从领域学者 h-index 值分布看，前十国家领域学者 h-index 值分布在 10 至 17 之间。中国学者 h-index 均值在领域学者数量前十的国家之中处于较低位置。

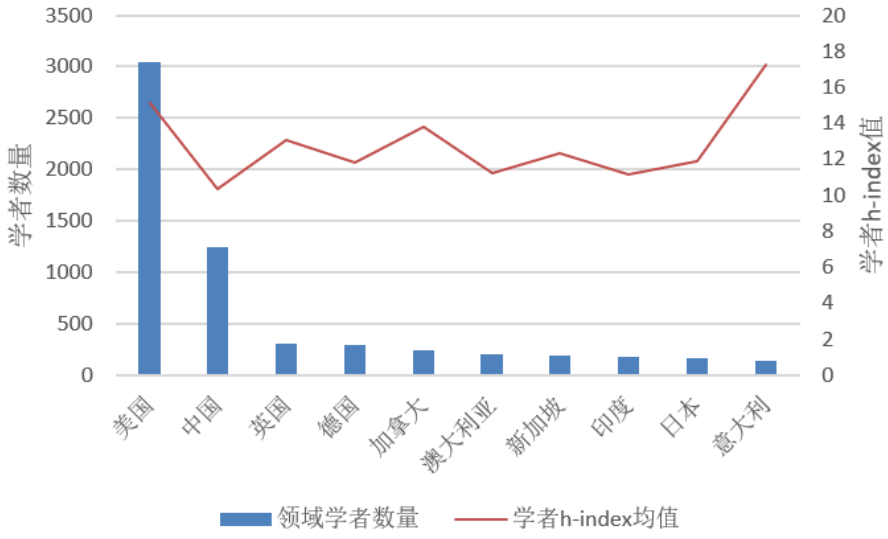


图 10 学术搜索领域学者数量 TOP10 的国家及该国学者 h-index 均值

美国领域学者的论文发表量最多，为 1483 篇；美国领域学者论文篇均被引用量为 49.5，仅处于加拿大学者论文的篇均被引用量，位列第 2。中国学者的领域论文数量和论文的篇均引用量整体上处于居中位置，与美国存在较大差距。

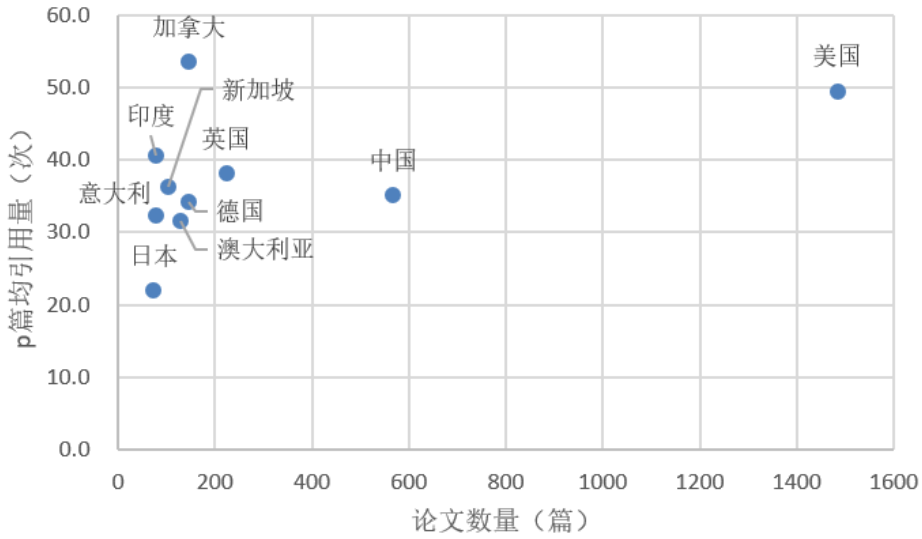


图 11 学术搜索领域学者数量 TOP10 国家学者的论文发表量和篇均引用量

3.1.4 学者机构分布

在领域学者数量排名前十的机构（如图 12 所示）之中，清华大学位列第一，领域学者数量为 231 人，但学者 h-index 均值在排名前十的机构中仍处于偏低位置。微软和卡内基梅隆大学分别位列第二位、第三位，领域学者数量分别为 202 人、199 人，这两家机构的领域学者 h-index 均值都处于较高水平。

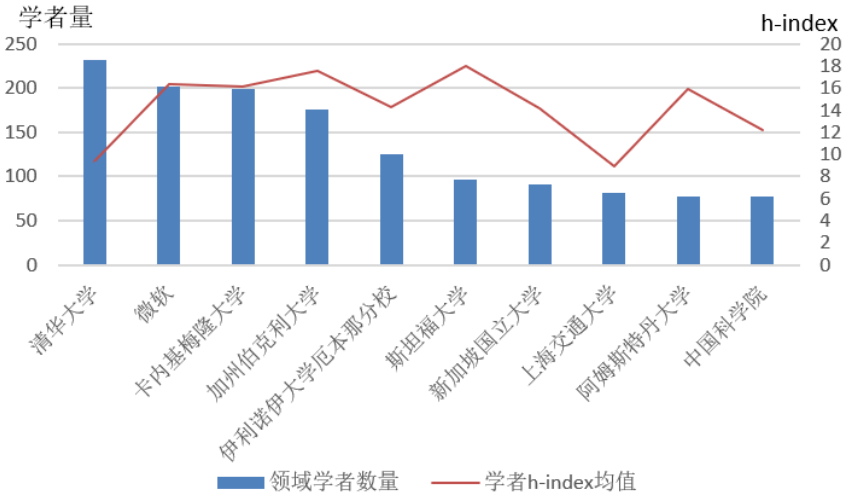


图 12 学术搜索领域学者数量 TOP10 机构及该机构学者 h-index 均值

清华大学领域学者的论文发表数量为 107 篇，领先于其他机构，但论文的篇均引用量在排名前十的机构中处于偏低位置。微软和卡内基梅隆大学领域学者的论文数量虽较多，但篇均引用量也处于中等水平。斯坦福大学领域学者虽然论文发表量不多，但篇均引用量则过百。

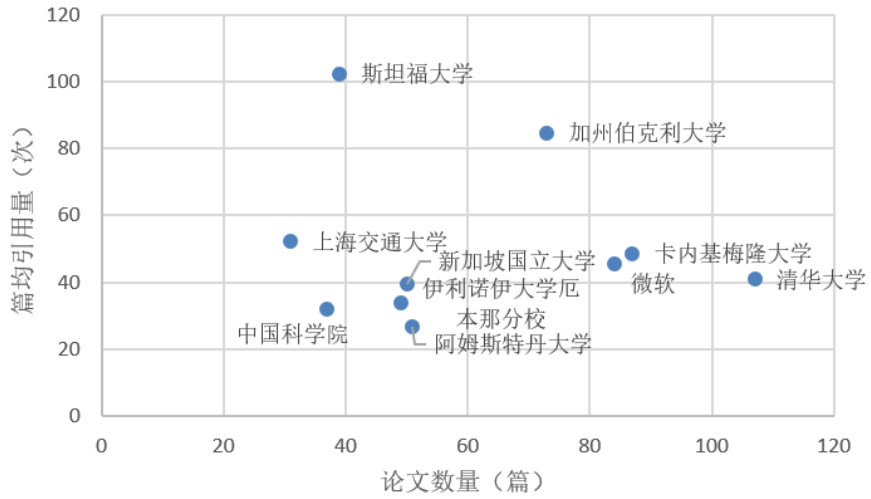


图 13 学术搜索领域学者数量 TOP10 机构的学者论文发表量和篇均引用量

3.1.5 领先机构学者研究重点

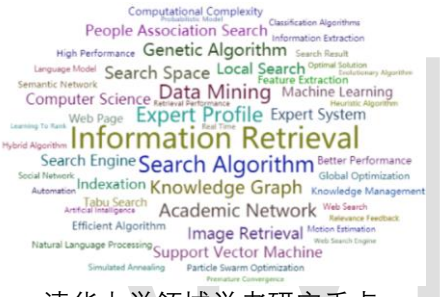
通过挖掘机构学者在学术搜索领域发表论文的主题及关键词情况，可以得出不同机构学者在该领域的重点研究方向。结果显示，清华大学、微软、卡耐基梅隆大学与加州伯克利大学的**领域学者研究重点都涵盖信息检索、信息检索系统等领域，**但**在细分方向上稍有差别：**清华大学和加州伯克利大学的学者更多研究搜索算法，微软学者更多研究语言模型和机器学习，卡耐基梅隆大学领域学者更多研究语言或视频搜索，加州伯克利大学学者还较多研究了机器翻译等相关方向。如图 14 所示。

领域学者数量最多的清华大学，其领域学者的重点研究方向是 Information Retrieval（信息检索）、Academic Network（学术网络）、Search Engine（搜索引擎）、Expert Profile（专家画像）、Knowledge Graph（知识图谱）以及 People Association Search（关联搜索）等。

微软公司领域学者的研究方向主要有：Information Retrieval（信息检索）、Search Engine（搜索引擎）、Language Model（语言模型）、Semantic Model（语义模型）和 Machine Learning（机器学习）等。

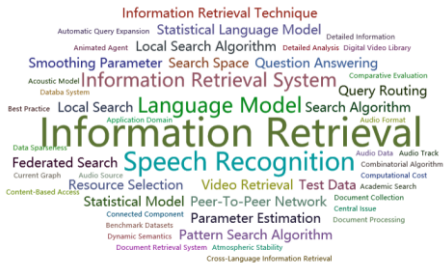
卡耐基梅隆大学领域学者的研究方向：Information Retrieval（信息检索）、Speech Recognition（语音识别）、Language Model（语言模型）、Video Retrieval（视频检索）和 Search Space（搜索空间）等。

加州伯克利大学领域学者的研究方向：Information Retrieval（信息检索）、Search Algorithm（搜索算法）、Private Information Retrieval（私人信息检索）、Data Mining（数据挖掘）、Machine Translation（机器翻译）、Probabilistic Model（概率模型）、Search Engine（搜索引擎）以及 Geographic Information Retrieval（地理信息检索）。



清华大学领域学者研究重点

微软领域学者研究重点



卡耐基梅隆大学领域学者研究重点

加州伯克利大学领域学者研究重点

图 14 机构领域学者的研究重点

3.1.6 学者跨机构合作情况

根据不同机构学者共同发表论文情况，可以得出领域跨机构研究合作状况。结果显示，学术搜索领域中国学者与其他国家学者合作发表论文数量 TOP10 之中（如图 15 所示），与美国合作的中国学者数量最多，共 325 人，合作的美国学者数量 288 人，**中美两国领域学者合作发表论文 153 篇，总被引频次 6,670**

次，远远领先于其他合作国家。其次是与新加坡合作的中国学者数量，共 79 人，参与合作的新加坡学者数量 61 人，两国领域学者合作发表论文 44 篇，总被引频次 1,931 次。合作国家中，中加两国学者合作论文量排在第三位，中国参与合作的学者共 58 人，加拿大参与合作的学者数量 25 人，两国领域学者合作发表论文 31 篇，总被引频次 885 次。

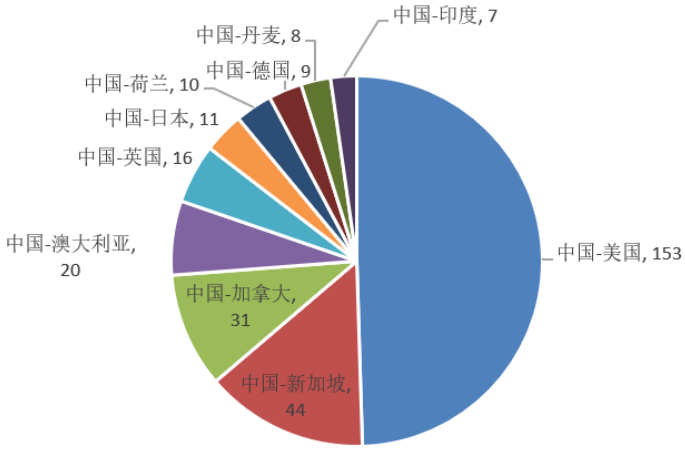


图 15 领域中国学者与其他国家学者合作发表论文情况（篇）

从与美国合作的中国学者数量趋势来看（如图 16 所示），**2009 年至 2019 年期间，与美国合作的中国领域学者数量波动较大**，中美合作领域学者量于 2016 年降至近十年的最低值，之后随着我国推行“一带一路”合作扩大对外开放，于 2017 年升至高峰，随后可能由于受到中美贸易战影响，参与合作的中国领域学者数量再次出现下滑。

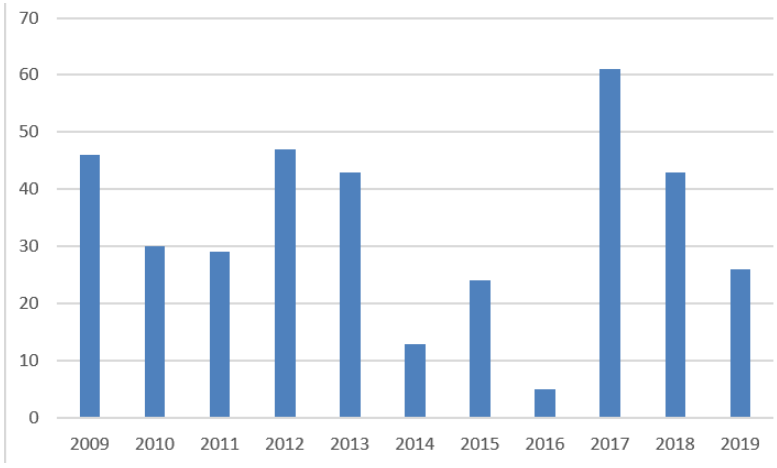


图 16 2009 年-2019 年与美国合作的中国领域学者数量

3.1.7 学者流动情况

通过学者发表论文时所在机构的变化，可以计算出学者的迁徙情况。从学者迁徙图中可以得出，2009 年至 2019 年，中国学者迁入迁出总人数呈现下降趋势，且迁入迁出人数基本持平。如图 17 所示。

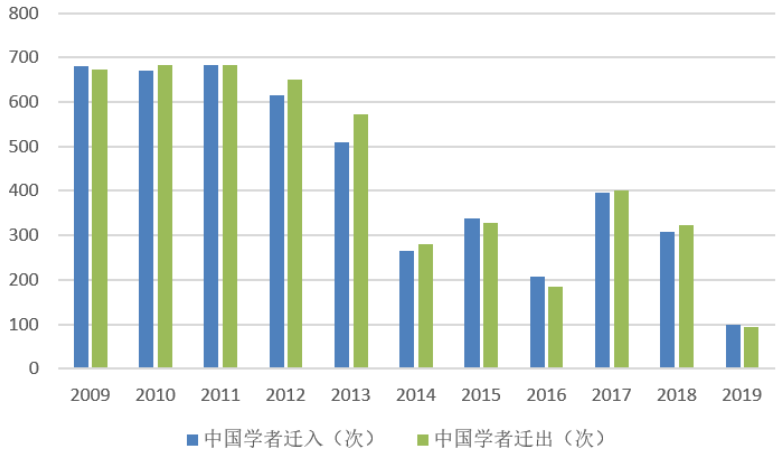


图 17 2009-2019 年期间学术搜索领域中国学者迁入迁出情况

全球范围内，领域学者迁徙人数排名前十的国家依次是美国、中国、英国、印度、德国、澳大利亚、加拿大、新加坡、韩国和日本，如图 18 所示。美国领域学者流动最大排名第一，迁徙总数为 12,686 人次，其中迁入 6,452 人次，迁出 6,234 人次。中国领域学者流动较大排名第二，迁徙总数为 9,644 人次，其中迁入 4,722 人次，迁出 4,872 人次。英国学者流动较为平缓，排名第三，迁徙总数为 1,980 人次，其中迁入 985 人次，迁出 995 人次。

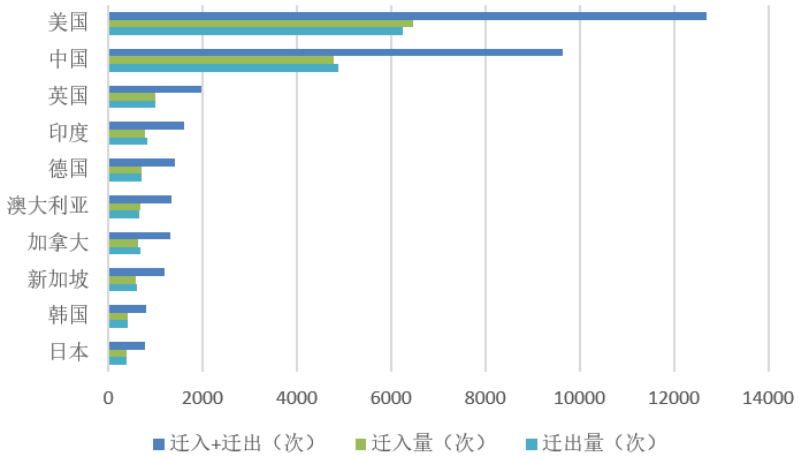


图 18 学术搜索领域学者迁徙总量 TOP10 国家

3.2 代表性领域学者介绍

通过 AMiner 人才智库的统计分析，可获得 AI 学术搜索领域学者信息，包括学者姓名、照片、职位、隶属机构、论文数量、论文引用量及研究领域等。本部分选取了领域内 h-index 较高的学者进行展示。

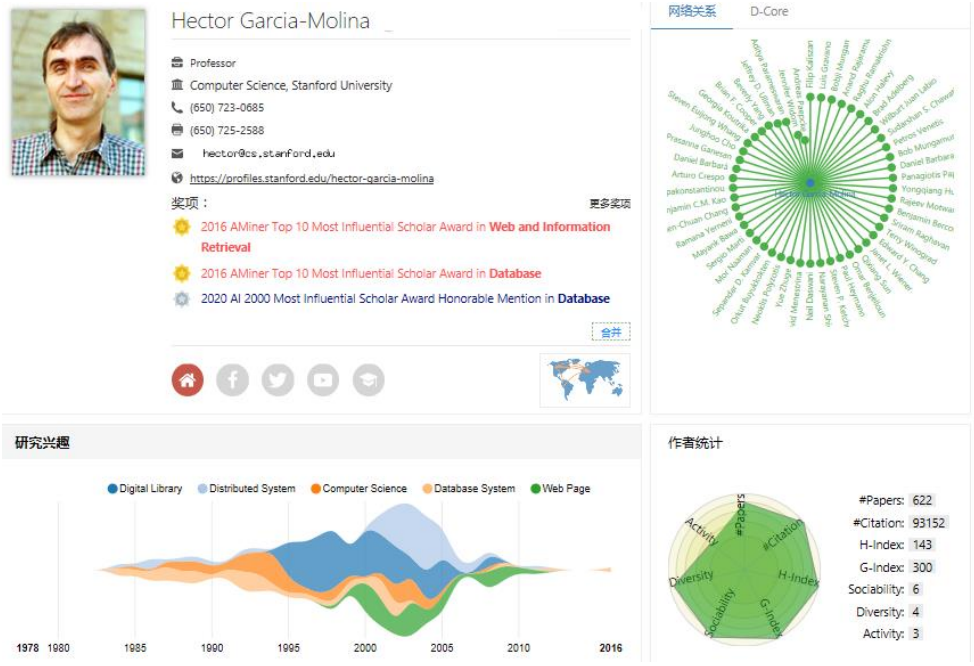
- **Hector Garcia-Molina**

现任斯坦福大学计算机科学系教授、ACM（the Association for Computing Machinery）和 AAAS（the American Academy of Arts and Sciences）Fellow，the National Academy of Engineering 成员。他的研究兴趣包括分布式计算系统、数字图书馆和数据库系统。

1975 年获得斯坦福大学电子工程硕士学位，1979 年获得计算机科学博士学位；2007 年拥有苏黎世联邦理工学院荣誉博士学位。

曾于 2001 年 1 月至 2004 年 12 月担任计算机科学系系主任；从 1997 年到 2001 年，他是 PITAC（the President's Information Technology Advisory Committee）的成员；从 1994 年 8 月到 1997 年 12 月，担任斯坦福大学计算机系统实验室主任；从 1979 年到 1991 年，在新泽西州普林斯顿大学计算机科学系任教。

1999 年荣获 ACM SIGMOD 创新奖。



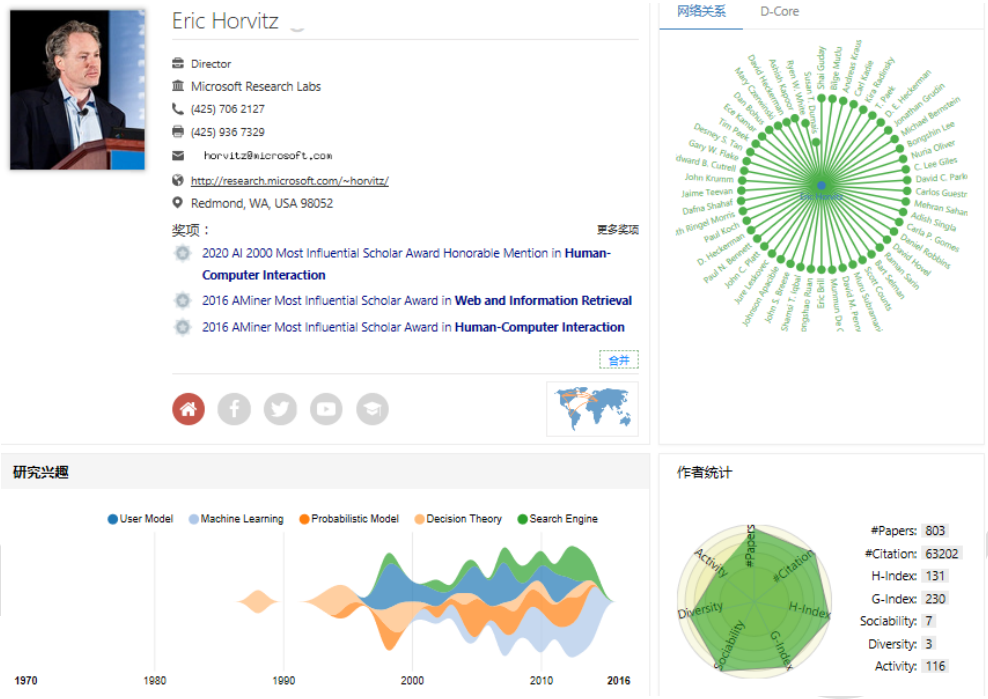
- **Eric Horvitz**

现任微软技术研究员兼首席科学官。

1993 年获得斯坦福大学博士学位。

曾任微软研究院主任、杰出科学家。

Horvitz 于 2020 年荣获 AI 2000 最具影响力学者奖人机交互荣誉奖。



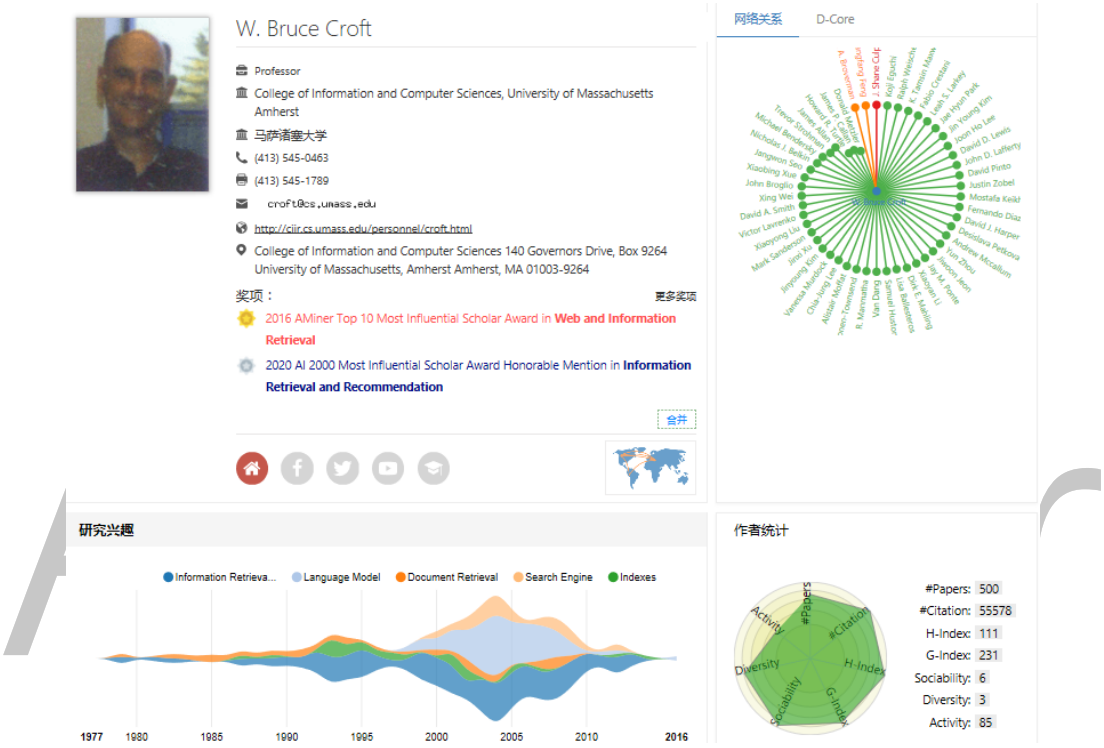
- **W. Bruce Croft**

现任马萨诸塞大学阿姆赫斯特分校计算机科学教授，也是智能信息检索中心的主任。研究兴趣在资讯检索的多个范畴，包括检索模式、表示法、网页搜寻、查询处理、跨语言检索及搜寻架构。

1979 年获得英国剑桥大学计算机科学博士，1974 年获得澳大利亚莫纳什大学计算机科学硕士，1973 年获得澳大利亚莫纳什大学计算机科学学士。

曾于 2015 年到 2017 年担任信息与计算机科学学院院长；2001 年至 2007 年担任马萨诸塞州立大学阿默斯特计算机科学系系主任。已经发表了 250 多篇网络信息抽取与推荐相关主题的论文。通过与 CMU 合作的 Lemur 项目，他的团队为学术界和工业界的许多人提供了开源搜索引擎和研究工具。

2000-2003 年曾为国家研究委员会计算机科学和电信委员会成员，1995-2002 年曾为 ACM Transactions on Information Systems 主编。在 1997 年当选为 ACM Fellow，在 2000 年获得 American Society for Information Science and Technology 研究奖，在 2003 年因其在信息检索研究的显著贡献而获得 ACM 信息检索（SIGIR）的终身成就奖。2013 年获得了 UKeiG Tony Kent Stix 奖，2014 年获得了 IEEE 计算机协会技术成就奖。



- **Rakesh Agrawal**

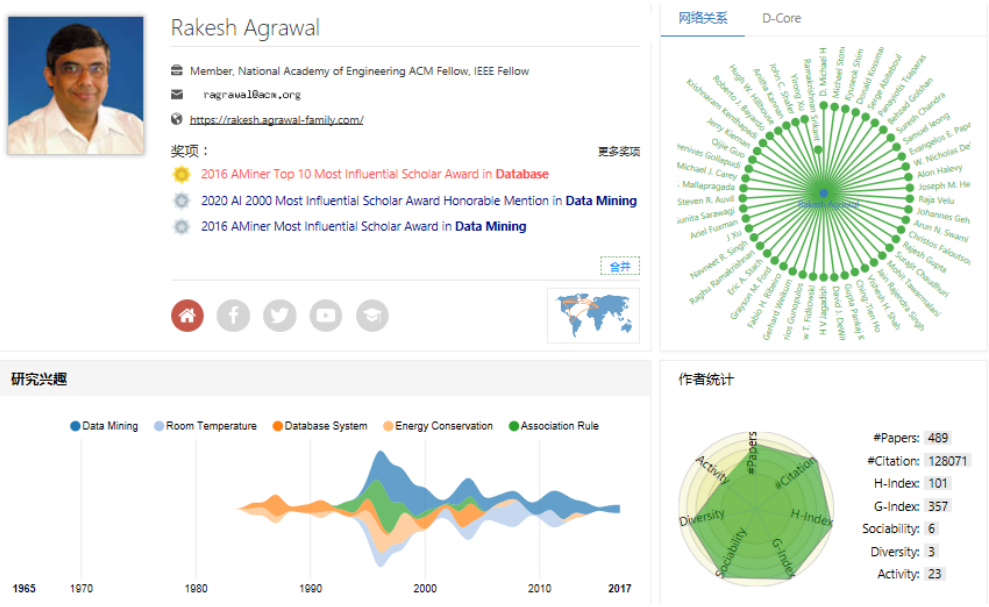
现为 National Academy of Engineering 成员、ACM Fellow 和 IEEE Fellow。

1983 年获得 Wisconsin-Madison 分校计算机科学哲学博士，1978 年获得国家工业工程研究所工业工程硕士，1975 年获得印度 Roorkee 电气电子和通信工程技术学院工程学士。

曾在 2006 年 3 月加入微软担任技术研究员，并领导微软研究院的搜索实验室。之前，是 IBM 的一名研究员，并在 IBM Almaden 研究中心领导 Quest 小组，以及在 Bell 实验室、印度最大的公司 Bharat Heavy Electricals Ltd.工作。

曾获得 ACM-SIGKDD 首届创新奖、ACM-SIGMOD Edgar F. Codd 创新奖、ACM-SIGMOD 时间测试奖（两次）、VLDB 10-Yr 最具影响力论文奖、ICDE 最具

影响力论文奖和计算机世界第一地平线奖的获得者。2003 年，《科学美国人》杂志将他列为 50 位顶尖科学家和技术专家之一。

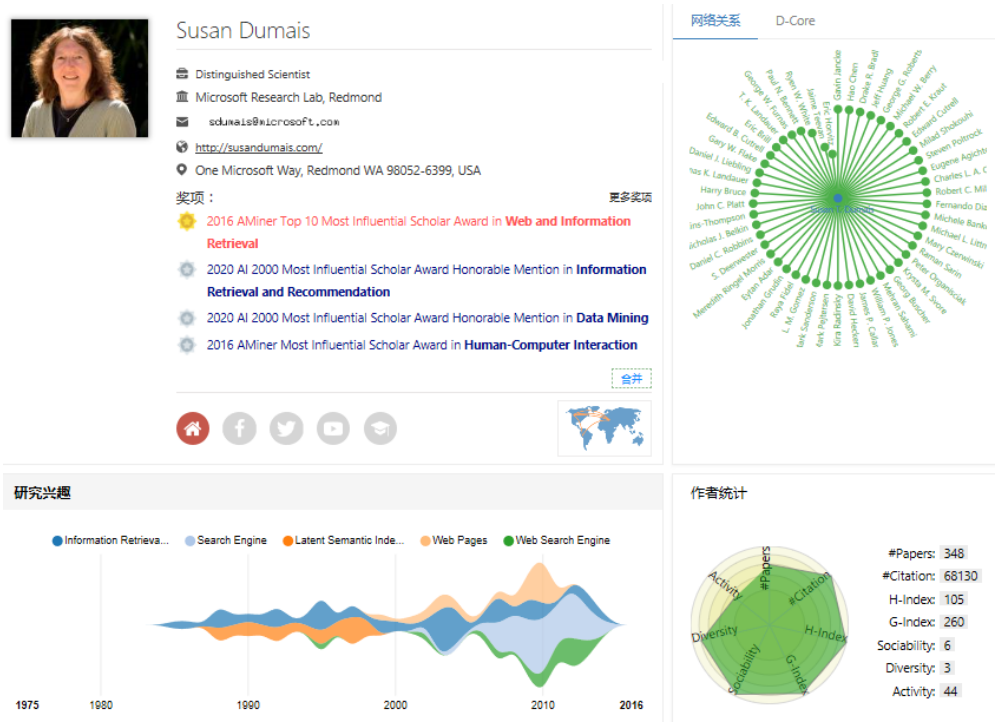


- **Susan Dumais**

现任微软研究实验室杰出科学家。研究领域包括信息检索改进算法和接口、人机交互。目前的研究主要集中在视线增强交互、信息系统的时间动态、用户建模和个性化、交互检索的新接口和搜索评估。之前，研究过个人信息管理、桌面搜索、问题回答、文本分类、协同过滤、用于改进搜索和导航的界面以及用户/任务建模。在搜索相关的创新方面，曾密切参与过必应、Windows 桌面搜索、SharePoint 门户服务器和 Office 在线帮助等几个微软小组。

毕业于印第安纳大学贝茨学院。

从 1997 年 7 月开始在微软研究院工作。



- **Weiying Ma 马维英**

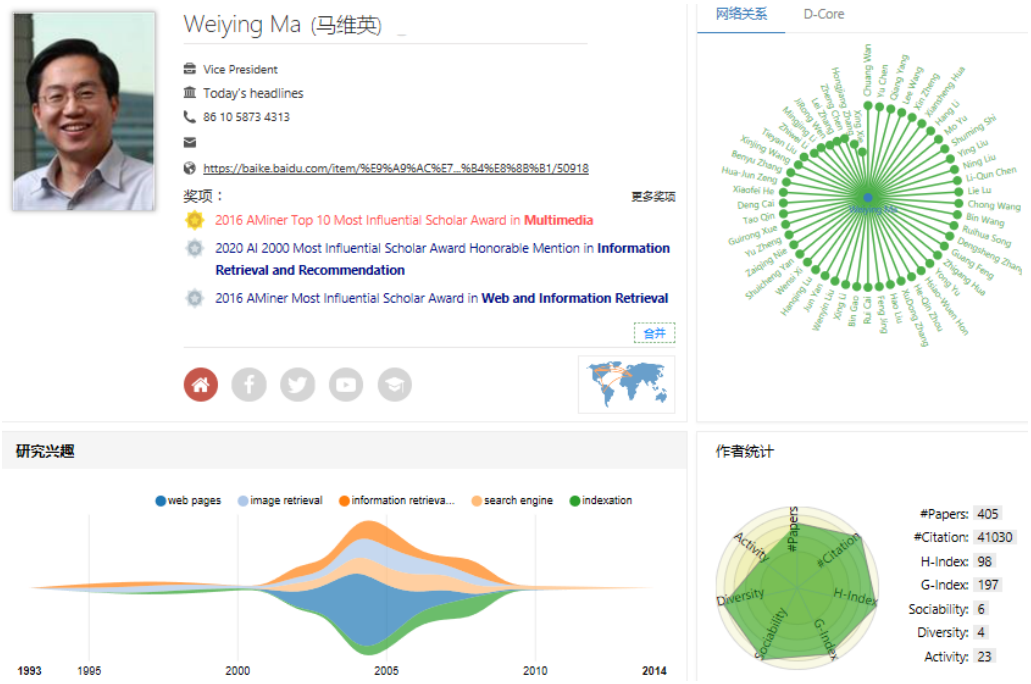
现任今日头条（Today's Headline）副总裁，负责人工智能实验室。是计算机科学与电子领域顶级科学家，也是 IEEE Fellow、ACM 杰出科学家。

1997 年获得加州大学圣巴巴拉分校 (UCSB) 电子与计算机工程系博士、1994 年获得该校电气与计算机工程系理学硕士、1990 年获得台湾国立清华大学电气工程系学士。

曾任职于硅谷惠普实验室。2001-2017 年曾任微软亚洲研究院常务副院长，负责信息检索、互联网搜索技术、移动信息浏览等方面技术的研究。许多研究成果都被应用到 Windows Live 图片搜索、移动搜索和学术搜索等服务中。2017 年，辞职微软亚洲研究院，加入今日头条。

曾任 ACM/Springer 多媒体系统期刊的编委以及 ACM Transactions on Information System（TOIS）期刊的副主编。他还是许多国际会议的组织和程序委员会成员，如 ACM Multimedia, ACM SIGIR, ACM CIKM, WWW, ICME, CVPR, SPIE Multimedia Storage and Archiving Systems , SPIE Multimedia Communication and Networking 等。他还是 International Multimedia Modeling Conference (MMM) 2005 和 International Conference on Image and Video Retrieval (CIVR) 2005 的大会联合主席。

已在世界级会议和学报上发表论文逾 270 篇，并拥有 160 多项技术专利。

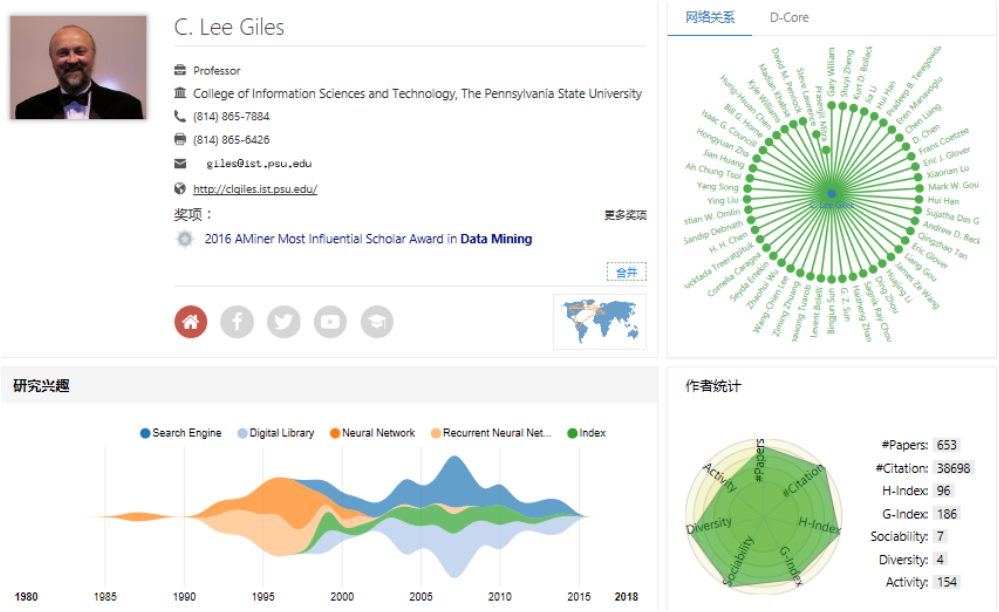


- **C. Lee Giles**

现任宾夕法尼亚州州立大学信息科学与技术学院教授，也是计算机科学与工程教授，供应链与信息系统教授，以及智能系统研究实验室主任，他是 CiteSeerX 项目的负责人，也是宾州州立大学 ChemXSeer 项目的共同负责人。他是 ACM 的 Fellow、IEEE Fellow、国际神经网络协会的 Fellow，AAAI 和 AAAS 的成员，也是 Sigma Xi, Tau Beta Pi 和 Eta Kappa Nu 的成员。

获得亚利桑那大学光学科学博士、密歇根大学物理学硕士、田纳西大学工程物理理学学士、田纳西州孟菲斯罗兹学院学士。

曾两次获得 IBM 杰出教员奖。曾在新泽西州普林斯顿的 NEC 研究所（现在的 NEC 实验室）担任高级研究员，在华盛顿特区空军科学研究办公室担任项目经理，在华盛顿特区海军研究实验室担任研究科学家。



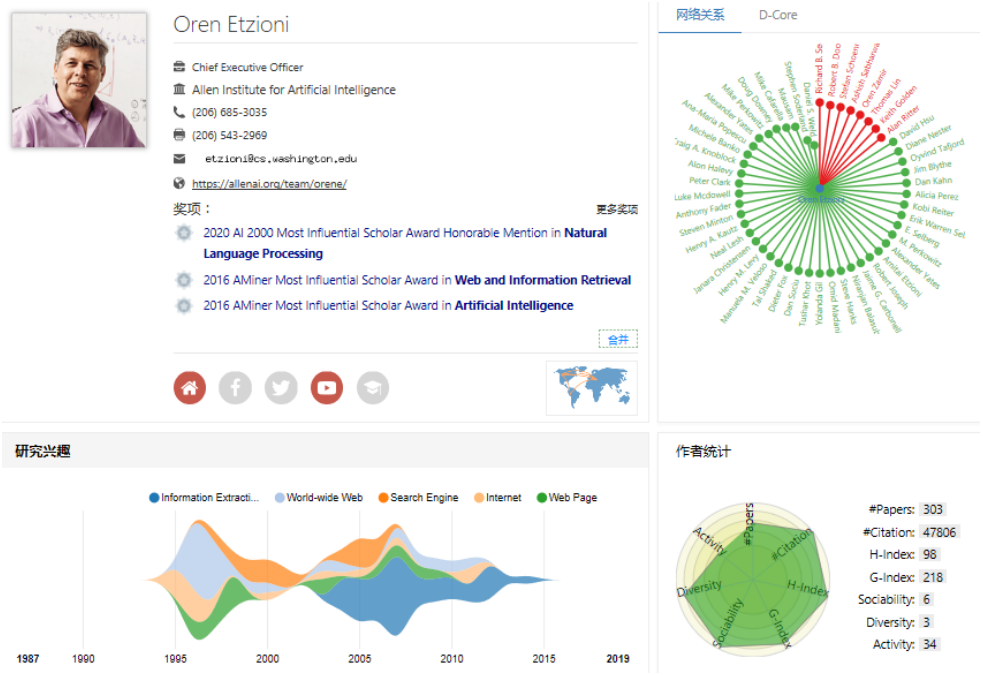
- **Oren Etzioni**

现任艾伦人工智能研究所的首席执行官、华盛顿大学的计算机科学系教授。Oren 的研究目标是解决人工智能的基本问题，特别是从文本中自动学习知识。

1991 年获得卡内基梅隆大学博士学位，1986 年获得哈佛大学学士学位。

曾是几家公司的创始人或联合创始人，其中包括 Farecast（2008 年卖给微软）和 Decide（2013 年卖给 eBay），他还撰写了 100 多篇技术论文，被引用次数超过 2.5 万次。

获得奖项包括 GeekWire's Hire of the Year (2014), Seattle's Geek of the Year (2013), the Robert Engelmore Memorial Award (2007), IJCAI 杰出论文奖 (2005), AAAI 研究员（2003）和国家年轻调查员奖（1993）。



- **James Allan**

现担任 UMass Amherst 大学信息与计算机科学学院的教授，也是智能信息检索中心（CIIR）的联合主任。工作重点领域包括交互式信息检索和组织，包括浏览和其他人机交互、自动信息组织、信息检索系统的评估、建立索引、检索和组织扫描书籍、理解搜索结果的争议性。

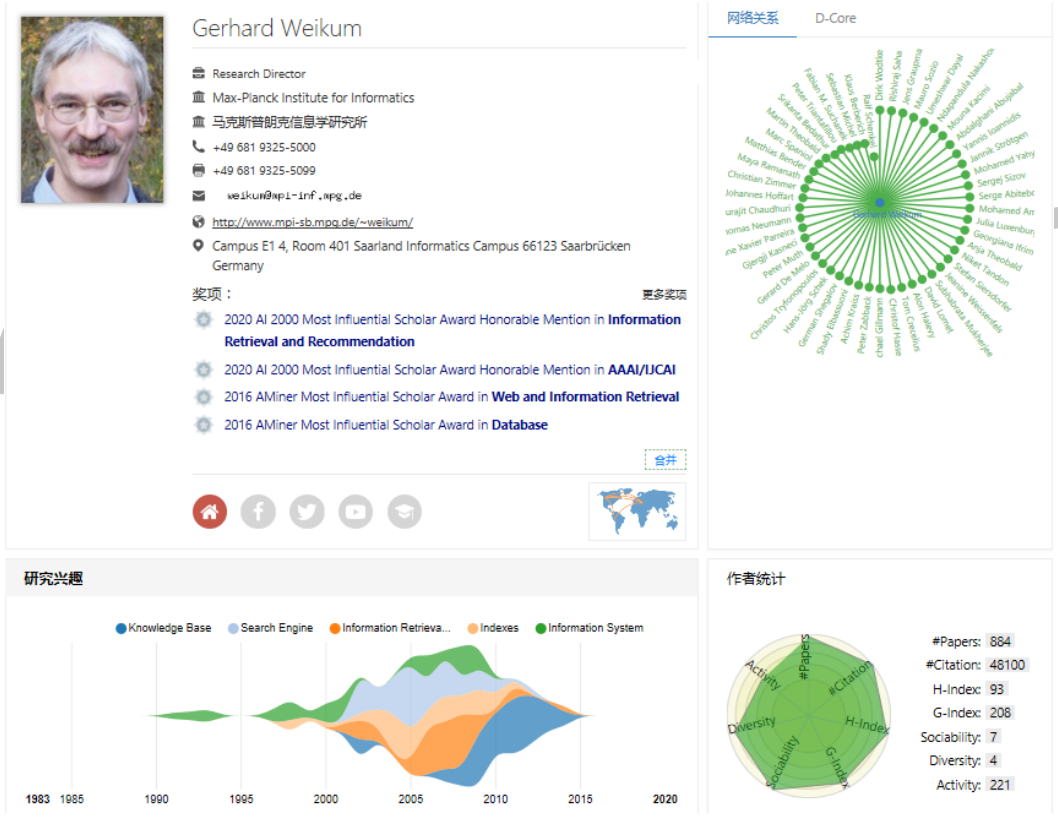
1995 年获康奈尔大学计算机科学博士学位，1991 年获康奈尔大学计算机科学硕士学位，1983 年在格林内尔学院获得数学学士学位。

- **Gerhard Weikum**

现担任德国 Saarbruecken 的 Max-Planck 信息学研究所(MPII)的研究主任，领导着该研究所的数据库和信息系统部门。他也是德国 Saarland University 计算机科学系的副教授，多模态计算和交互卓越集群的首席研究员。研究方向为分布式信息系统数据库性能优化（自动调优）和自组织（自主计算）数据库与 IR 集成(数据库系统与信息检索)、信息抽取和知识获取。

获得德国 Darmstadt 大学博士学位。

曾在德国 Saarbruecken 的 Saarland 大学、瑞士苏黎世联邦理工学院、美国德克萨斯州奥斯汀的 MCC 任职，并在华盛顿雷德蒙德的微软研究院担任访问高级研究员。



4

产品篇

学术搜索产品层出不穷。除了政府部门、大学、图书馆等机构，一些著名搜索引擎公司也在致力于相关搜索产品开发。随着新兴技术发展，搜索产品除了传统的文献检索功能以外，一些嵌入 AI 技术特色的学术搜索也陆续涌现。

4.1 学术搜索产品的时间演化图

自上世纪 90 年代以来，很多机构都推出了功能不同、搜索范围各异的学术搜索引擎。这些机构来自中国、美国、荷兰、加拿大、英国以及德国等多个国家，其中既有高校图书馆，也有政府部门或非营利组织，但多数是商业公司。

初时，一些出版机构通过数字化整合自身资源而推出学术搜索产品，例如，荷兰老牌出版商爱思唯尔早于 1997 年推出的 ScienceDirect 以及在同一年原汤森路透公司推出的 Web of Science 产品，中国知网于 1999 年推出类似的搜索产品。

如图 19 所示，2004 年，有多个学术搜索产品问世，其中以谷歌学术最为知名，也达到传统学术搜索有史以来的技术高峰。之后几年，AMiner、微软学术、百度学术、Semantic Scholar 等商业产品相继上线。同时，AI 技术被大规模地应用于学术搜索领域，市场上各个搜索产品不断升级、增添各自 AI 特色。与其他产品通过升级迭代引入 AI 技术不同，Semantic Scholar 是以 AI 理念创新设计而推出的学术搜索产品，其“智能”功能更为显著。

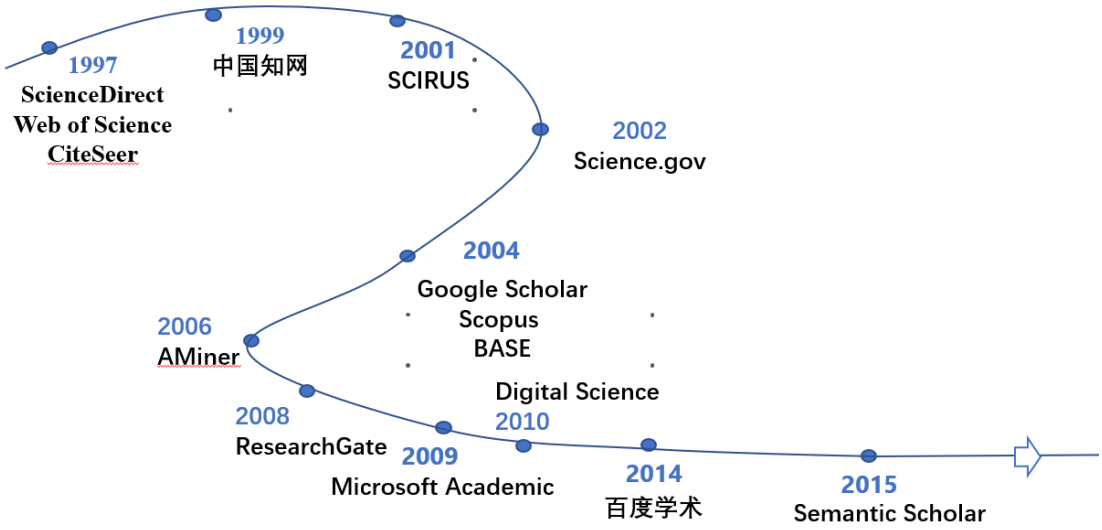


图 19 学术搜索产品的时间演化图

4.2 产品分类

根据资源开放程度，学术搜索产品分为融合开放数据和服务、专门数据库和服务。融合开放数据和服务的产品实现多源异构数据接入汇聚，通常可以公开访问，个别产品需要用户先进行注册才可以访问；对于文献搜索结果，提供免费全文或者提供获取方式链接以便用户单独购买。专门数据库和服务则是集成特定资源的数据库平台，一般不支持互联网公开访问，通常需要用户付费或获得权限之后才能访问使用。具体如表 1 所示。

表 1 基于资源开放程度的学术搜索产品分类

分类	融合开放数据和服务	专门数据库和服务
特征	支持互联网公开访问	不支持互联网公开访问
代表产品	Google Scholar Microsoft Academic BASE CORE Science.gov Semantic Scholar Baidu Scholar AMiner	Web of Science ScienceDirect Scopus 中国知网 CNKI

根据搜索数据库的学科覆盖范围，学术搜索产品又可以分为综合性学术搜索产品和垂直学术搜索产品，如表 2 所示。目前，以覆盖多个不同学科资源为特征的综合性学术搜索产品占据主流。

表 2 基于覆盖学科的学术搜索产品分类

分类	综合性学术搜索	垂直学术搜索
特征	多个不同学科	专门学科
代表产品	Google Scholar Microsoft Academic BASE CORE Science.gov Semantic Scholar Baidu Scholar Web of Science ScienceDirect Scopus 中国知网 CNKI	CiteSeer（计算机） DBLP（计算机）

4.3 主要产品一览

目前，主要的国内外学术搜索产品如下表所示。可以看出，多数产品以多学科文献集成数据库为主要特征，少数产品则是以 AI 技术驱动学术搜索的定位而问世，例如 AMiner、Semantic Scholar。

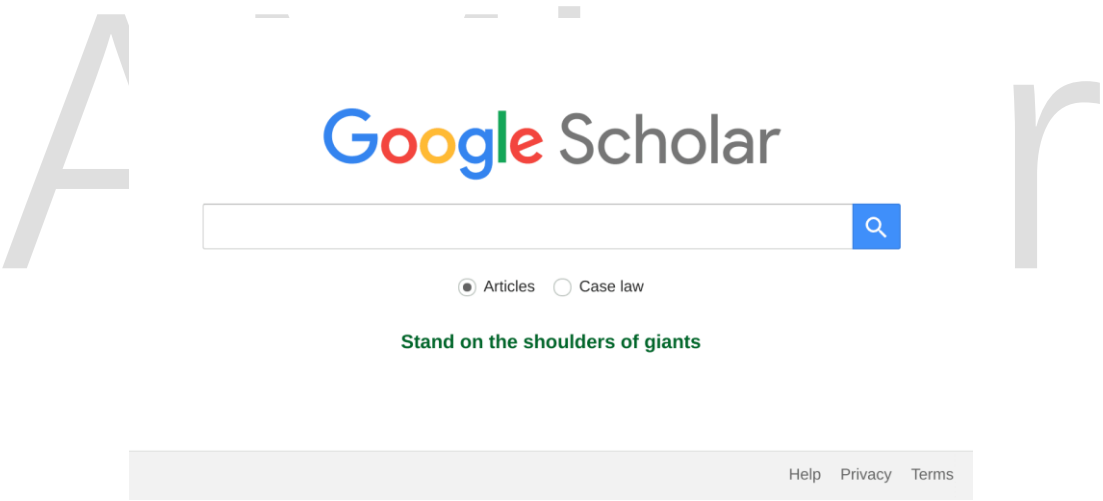
表 3 主要学术搜索产品

公司	产品名称	发布年份	特性描述
英国 Open 大学	**CORE**	-	汇总全球开放访问的研究论文
特里尔大学	**DBLP**	-	计算机类英文文献的集成数据库系统
汤森路透	**Web of Science**	1997	大型的综合性、多学科的核心期刊引文索引数据库。科学引用指标、交叉学科引用指标
爱思唯尔	**ScienceDirect**	1997	多学科全文数据库
宾夕法尼亚州立大学	**CiteSeerX**	1997	专注计算机和信息科学领域的文献检索平台
中国知网	**CNKI**	1999	提供各类资源统一检索、统一导航、在线阅读和下载服务
美国能源部	**Science.gov**	2002	美国科学信息门户
谷歌	**Google Scholar**	2004	大规模学术论文索引与快速检索
爱思唯尔	**Scopus**	2004	全学科、交叉学科文献收集
比勒费尔德大学	**BASE**	2004	多学科的学术搜索引擎
清华大学	**AMiner**	2006	AI 赋能学术搜索
ResearchGate	**ResearchGate**	2008	科研社交网络服务
Nature	**Digital Science**	2010	为科学研究过程提供数据服务和工作流解决方案
微软	**Academic Search**	2014	学术搜索
百度	**百度学术**	2014	提供海量中英文文献检索的学术资源搜索平台
AI2	**Semantic Scholar**	2015	AI 驱动的研究工具

4.3.1 谷歌学术 Google Scholar

Google Scholar (http://scholar.google.com)是 Google 公司于 2004 年 11 月发布的一个跨学科的、免费学术搜索引擎。已陆续开发出英文、中文、丹麦、芬兰、荷兰、挪威、瑞典等多个语言版本，其中，2006 年 1 月扩展到中文学术文献领域。2012 年推出谷歌学术计量（GSM），用来评价各个领域杂志的影响力。谷歌学术计量系统主要包括 h 指数、h 核心（h-core）、h 中值（h-median）等。

Google Scholar 是特别为学术检索而推出的网络应用，滤掉了普通搜索结果中大量的垃圾信息，特点是能检索阅读的学术资料极其丰富，提供引用信息，便于用户组合选择的结果排列方式等。搜索结果显示的是内容与搜索词存在相关性的文章，排名则与该文引用增加的速度有关，并在检索结果中实现了对不同数据源中相同资源的不同版本的归并和显示功能，并提供详细显示这些版本的链接。



4.3.2 微软学术/必应学术 Microsoft Academic

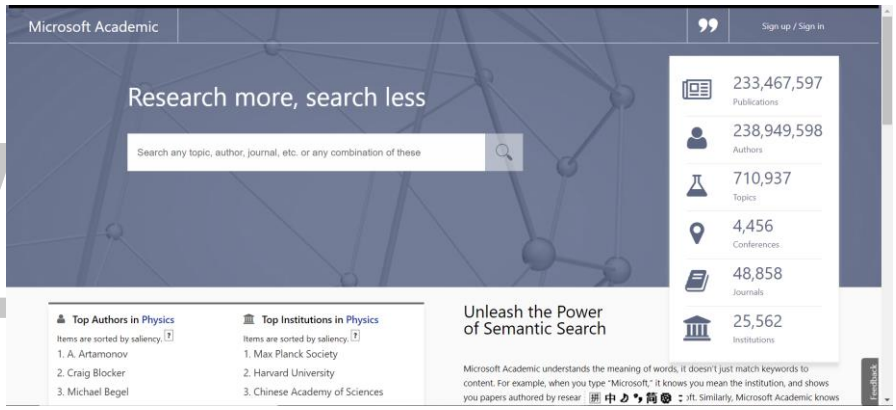
2009 年 11 月，微软亚洲研究院启动了一项新的专门研究科学信息的网络搜索服务，即微软学术（https://academic.microsoft.com/home）。微软学术搜索（Microsoft Academic，简称 MA）致力于构建学术资源平台，基于必应的大数据搜索技术及微软亚洲研究院的先进算法，同时整合了 Azure 云计算能力，为学术用户提供优质的、全球的多语种文献检索服务。

MA 为研究员、学生、图书馆馆员和其他用户提供了一个智能、新颖的搜索平台，方便用户查找学术论文、知名学者、国际会议、权威期刊等信息。同时，

它作为一个研究试验平台，展现了研究院在对象级别垂直搜索、命名实体的抽取和消歧、数据可视化等研究领域的最新研究成果。

MA 是一个免费的语义搜索引擎，而不是基于关键字的搜索引擎。微软学术搜索利用机器学习、语义推理和知识发现方面的技术进步，帮助用户探索学术信息。MA 是一款实现按领域检索的学术搜索引擎，包括了 15 个领域，每个领域包含若干学科方向。它是上一个版本“Microsoft Academic Search”优化后的成果。Microsoft Academic Search 曾经通过 Windows Phone Client 推出过移动学术信息检索服务功能，用户可以使用手机进行学术文献检索，目前已下架。

目前，微软学术搜索已嵌入必应搜索引擎，在国内被嵌入到必应学术搜索（https://cn.bing.com/academic）。



MA 支持以下 **6 种查询实体类型：**

- ✓ **作者-**出版物的个人作者。
- ✓ **机构-**作者机构是作者在发表论文时所属的机构。
- ✓ **论文-**出版物标题。
- ✓ **期刊-**学术期刊的名称。
- ✓ **主题**-研究领域，由出版商关键词和微软学术算法确定。
- ✓ **会议**-展示研究成果的场所。

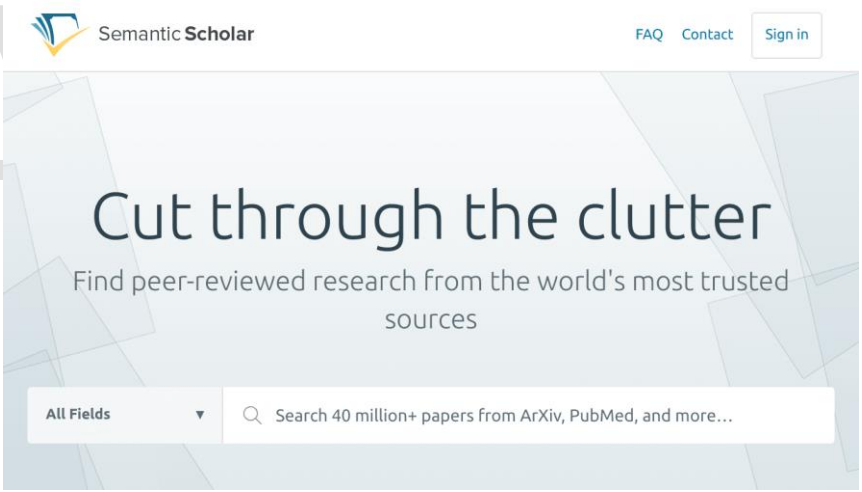
4.3.3 语义学术 Semantic Scholar

微软联合创始人保罗•艾伦于 2014 年出资成立艾伦人工智能研究所（Allen Institute for Artificial Intelligence，简称 AI2），致力于进行人工智能和计算机科学研究，其总部位于西雅图。2015 年 11 月 2 日，AI2 研究所推出的一款基于人

工智能的、全新的、免费的学术搜索引擎 Sematic Scholar（https://www.semanticscholar.org/）。

Sematic Scholar 从 PubMed、Nature、ArXiv 等专业期刊和 ScientificAmerican、WIRED、Discover 等专业媒体抓取文献和科学报道，并且使用 AI 技术分析作者、引用、主题、图形信息，服务于领先的科学研究。它利用机器学习技术从论文中抽取意义和识别联系，然后将这些见解呈现出来，帮助学者快速获得深入的理解；致力于使用人工智能算法发现研究主题之间隐藏关联，以提供更具相关性和影响力的搜索结果。

以往用于搜索引擎的“人工智能”表现在基于网络蜘蛛的智能化信息抓取、基于语义技术的用户意图自动识别及个性化搜索等，而 **Semantic Scholar 则基于深度学习而实现系统对论文内容的理解**，更接近目前所实现的人机大战模式的人工智能，将更有利于帮助用户筛选有用信息，提高学术信息搜索和过滤的效率。



4.3.4 百度学术 Baidu Xueshu

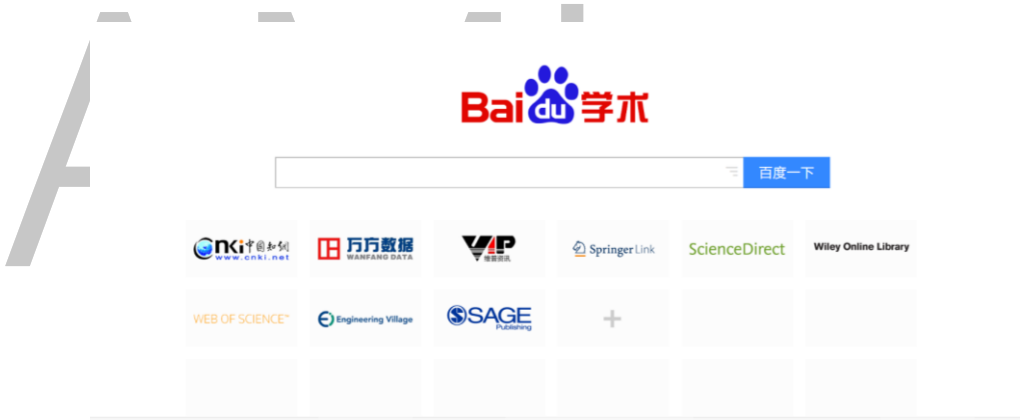
百度学术（http://xueshu.baidu.com）于 2014 年 6 月上线，是百度旗下的免费学术资源搜索平台，致力于将资源检索技术和大数据挖掘分析能力贡献于学术研究。自成立以来，百度学术已推出了研究点分析、相关热搜词分析，具有深入计量文献的内容特征。

百度学术搜索引擎的 UI 设计与功能设计上很大程度借鉴了 Google Scholar 的思路，为用户提供了个人学术管理和可视化功能，包括研究热点分析可视化和学术成果可视化等。虽然是中文界面，百度学术的索引结果包括中英文论文。

百度学术目前提供以下两大类服务：

1. 学术搜索：支持用户进行文献、期刊、学者三类内容的检索，并支持高校和科研机构图书馆定制版学术搜索。搜索结果列表中可呈现文献部分摘要。虽然大部分文献无法下载全文，但是提供了多个下载来源（包括免费来源）链接。用户可以通过选择所需下载源来获得文献全文。

2. 学术服务：支持用户“订阅”感兴趣的关键词、“收藏”有价值的文献、对所研究的方向做“开题分析”、进行毕业论文“查重”、通过“单篇购买”或者“文献互助”的方式获取所需文献、在首页设置常用数据库方便直接访问。



4.3.5 AMiner

AMiner 学术搜索是由清华大学计算机系教授唐杰率领团队研发，于 2006 年正式上线。基于文献、专利、成果和专家信息深入分析挖掘，AMiner 科研智能搜索引擎构建了专家画像和知识图谱，挖掘知识推理网络。平台以**科研人员、科技文献、学术活动**三大类数据为基础，构建三者之间的关联关系，深入分析挖掘，面向全球科研机构及相关工作人员，提供学者、论文文献等学术信息资源检索以及面向科技文献、专利和科技新闻的语义搜索、语义分析、成果评价等知识服务。典型的知识服务包括：学者档案管理及分析挖掘、专家学者搜索及推荐、技术发展趋势分析、全球学者分布地图、全球学者迁徙图、开放平台等。

该平台利用数据挖掘和社会网络分析与挖掘技术，提供研究者语义信息抽取、面向话题的专家搜索、权威机构搜索、话题发现和趋势分析、基于话题的社会影响力分析、研究者社会网络关系识别等功能。

4.3.6 BASE

BASE（http://www.base-search.net/）是德国比勒费尔德（Bielefeld）大学图书馆开发的一个多学科的学术搜索引擎，使用 Solr/Lucene 的开源搜索技术（2011 年 5 月之前采用的是微软 FAST 搜索和传递技术），提供对全球异构学术资源的集成检索服务。它整合了德国比勒费尔德大学图书馆的图书馆目录和大约 160 个开放资源（超过 200 万个文档）的数据。BASE 项目于 2003 年启动，2004 年“范本软件”公开发行。

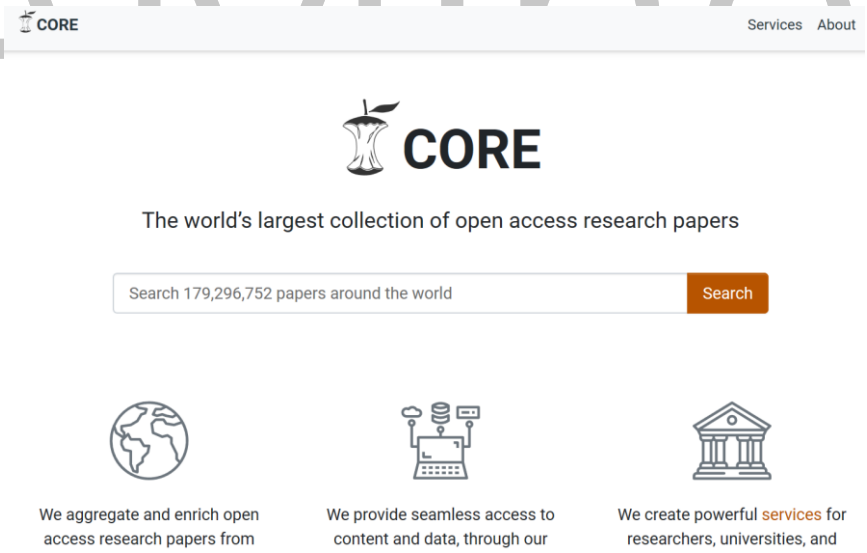
目前，BASE 已经注册成为 OAI 服务提供者。BASE 提供英、德、法、中等 8 种语言检索，默认语言是德语。

BASE 的最大特色是可以进行精化检索。系统分别从作者、主题、来源、文档大小、出版日期、文档类型、语种、文件格式 8 个方面对检索结果进行分析，按上述条件分别析取出检索结果中包含的知识元，以析出的知识元为依据对命中结果进行统计，将统计结果显示给用户。例如，对作者进行分析，首先从命中结果中析取出各条资源记录的作者，然后对命中结果中每位作者所占有的作品数量进行统计，将作者名称与结果中包含的该作者作品数量形成统计表返回给用户，使用户可以清楚地了解检索结果按作者的分布情况，并可以通过链接显示结果集中该作者的所有资源。

4.3.7 CORE

CORE（https://core.ac.uk/）是由英国开放大学和 Jisc 提供的非营利服务。它收集来自世界各地存储库和期刊的所有开放存取研究成果，并将其提供给公众。

CORE 从世界各地的数据提供商那里收集研究论文，包括机构和主题存储库、开放存取和混合期刊，只提供通过开放获取资源的文献全文，支持访问原始数据，同时，支持发现、推荐以及管理内容。



4.3.8 Science.gov

Science.gov（https://www.science.gov/）是 2002 年推出的美国科学信息门户网站，由美国能源部（DOE）主办。该网站提供免费获取研究和开发的成果以及来自 13 个联邦机构的科学组织的科学和技术信息。Science.gov 旨在为科学家

和工程师，图书馆和商业社区，学生，教师，企业家以及任何对科学感兴趣的人提供免费服务。用户无需注册即可使用。

Science.gov 检索结果包括政府机构网站以及相关的数据库。

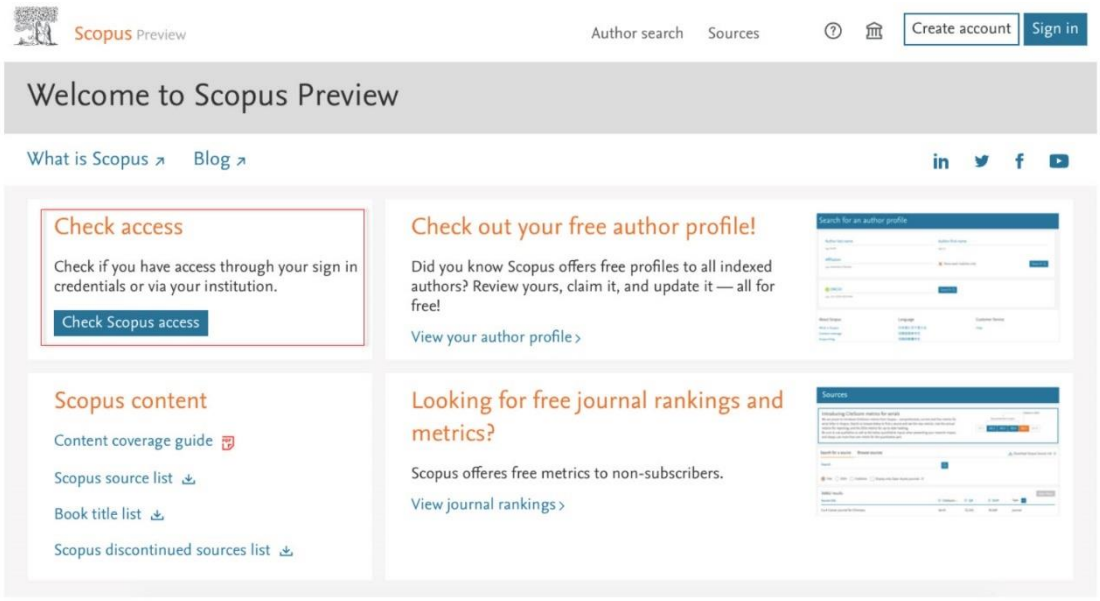
Science.gov 是由跨机构 Science.gov 联盟管理，该联盟包括来自以下联邦机构的科学和技术信息组织的代表：农业部、商务部、国防部、教育部、能源部、卫生与人类服务部、国土安全部、交通运输部、环保局、政府出版社、国家航空和航天局、国家科学基金会，这些机构约占联邦研发预算的 97%。

4.3.9 Scopus

Scopus（https://www.scopus.com/）是规模最大的同行评议文献（科学期刊、书籍和会议记录）的摘要和引文数据库，由荷兰出版商爱思唯尔（Elsevier）于 2004 年 11 月正式推出。Scopus 提供全球科学、技术、医学、社会科学、艺术和人文等领域研究成果的全面概述，并提供跟踪、分析和可视化研究的智能工具。Scopus 在全球有超过 3,000 家学术、政府和企业机构用户，并且是支持 Elsevier Research Intelligence 资源的主要数据来源。

Scopus 由全球 21 家研究机构和超过 300 名科学家共同设计开发而成的。

Scopus 荣获 2005 年“全球信息产业最佳科学技术医学信息产品奖”。



4.3.10 ScienceDirect

ScienceDirect（https://www.sciencedirect.com/）是荷兰全球著名的学术期刊出版商爱思唯尔（Elsevier）从 1997 年开始推出的一个全文数据库，又称为 ScienceDirect OnSite（Elsevier 电子期刊全文），致力于将该公司的全部印刷版期刊转换为电子版。

ScienceDirect 内容覆盖爱思唯尔旗下各品牌及合作机构的 2,500 多种期刊和超过 35,000 种图书。ScienceDirect 全文数据库涉及众多学科：计算机科学、工程技术、能源科学、环境科学、材料科学、数学、物理、化学、天文学、医学、生命科学、商业、及经济管理、社会科学等。

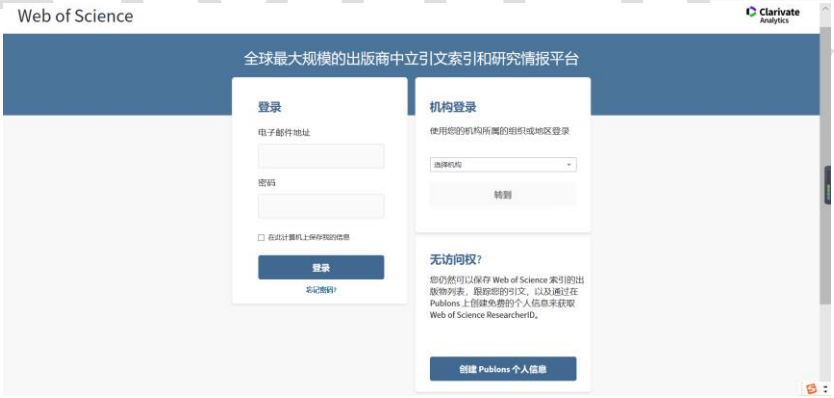
4.3.11 Web of Science

Web of Science（http://webofscience.com/）是一个大型的综合性、多学科的核心期刊引文索引数据库。最初是由美国科学信息研究所制作，1997 年 Thomson 公司将 SCI（Science Citation Index，创立于 1963 年）、SSCI（Social Science Citation Index，创立于 1973 年）以及 AHCI（Arts & Humanities Citaion

Index，创立于 1978 年）整合，利用互联网开放环境，创建了网络版的多学科文献数据库 Web of Science。目前由科睿唯安（Clarivate Analytics）维护。

Web of Science 包括三大引文数据库(科学引文索引(Science Citation Index，简称 SCI）、社会科学引文索引（Social Sciences Citation Index，简称 SSCI）和艺术与人文科学引文索引（Arts & Humanities Citation Index，简称 A&HCI））和两个化学信息事实型数据库(Current Chemical Reactions，简称 CCR，和 Index Chemicus，简称 IC)，以及科学引文检索扩展版 (Science Citation Index Expanded，简称 SCIE）、科技会议文献引文索引（Conference Proceedings Citation Index-Science，简称 CPCI-S）和社会科学以及人文科学会议文献引文索引（Conference Proceedings Citation index-Social Science&Humanalities，简称 CPCI-SSH）三个引文数据库，以 ISI Web of Knowledge 作为检索平台。

内容涵盖自然科学、工程技术、生物医学、社会科学、艺术与人文等领域，最早回朔至 1900 年。该数据库文献收录全面，检索界面友好，是业界标准（影响因子即是以此网站的数据为准）。



4.3.12 中国知网

中国知网（https://www.cnki.net/）是一个综合性数据库，于 1999 年 3 月，由王明亮提出建设，并被列为清华大学重点项目。知网全称是国家知识基础设施（China National Knowledge Infrastructure，简称 CNKI）。

知网收录了各学科期刊论文、图书、学位论文、报纸、会议论文、专利、标准等多种类型的文献。其检索模式多，且提供多种链接，便于查找相关文献。有

结果分组分析功能及多种排序方式，方便选择文献，并提供文献被引用情况。其搜索到的文献结果的下载格式有两种：CAJ 格式和 PDF 格式。

2017 年 11 月，知网首页升级，划分为文献、知识元、引文三大检索入口，推出中英文跨库智能检索。

4.4 产品覆盖的学术资源

作为一个细分的垂直搜索，学术搜索产品的价值主要在于其占有的学术资源。目前，多数产品的学术资源均是覆盖多个学科的期刊或会议论文、学位论文、图书、报告等文献。有的产品还包括专利数据，例如 Google Scholar 和百度学术。主要产品的资源覆盖详情见表 4。

表 4 主要学术搜索产品的资源覆盖情况

产品名称	覆盖学科	覆盖资源
Academic Search	全学科	约 2.3 亿论文，2.4 亿位作者，71 万个主题，以及 4,451 个会议，48,840 个期刊，2.6 万家机构。
AMiner	全学科	覆盖全球 1.36 亿学者、2.3 亿篇论文、7.5 亿论文引用关系、879 万知识概念，以及超过 160 个特色专家子库
BASE	全学科	来自期刊文章、预印本、数字收藏、图像/视频或研究数据等 7,870 个资源的约 1.6 亿文档
CiteSeerX	计算机、数学	超过 600 万份文档，其中有近 600 万独特作者和 1.2 亿次引用。
CNKI	全学科	未说明
CORE	Open Access	世界各地 9,897 家数据提供商的 1.8 亿篇开放存取文章
DBLP	计算机	500 万篇期刊会议文章、作者 250 万、会议 5,201 个，期刊 1,690 种
Digital Science	全学科	10 亿文献、9 亿引用文献和 2 千万作者

Google Scholar	全学科	论文约 2 亿，技术报告、书籍、预印本、专利等
ResearchGate	全学科	约 1.3 亿文献、1,500 万以上的学者数据
Science.gov	全学科	60 多个数据库，2,200 多个网站和超过 2 亿页的权威联邦科学信息，包括全文文档，引文，支持联邦资助研究的科学数据和多媒体
ScienceDirect	全学科	Elsevier Science 的 1,263 种全文电子期刊
Scopus	全学科	全球 5,000 多家出版商的 24,000 多种期刊，内容涉及人文、科学、技术及医学等
Semantic Scholar	全学科	期刊、学术会议资料或者是学术机构的文献。覆盖 19 个学科 1.8 亿论文。目前能检索到约 80%的免费论文文献。
Web of Science	全学科	254 个学科的 20,900 多种高影响力学术期刊
百度学术	全学科	期刊会议论文、专利、图书等 4 亿多篇学术文献、300 个学科研究方向、400 多万个中国学者主页、120 多万个国内外学术站点

- **Google Scholar 学术资源**

Google Scholar 覆盖资源由期刊文章全文、技术报告、预印本、论文、书籍和其他文档组成。包括约 2 亿篇论文，以及专利。它通过知识库、出版商平台和个人网页收集可在网上获得的科学论文，而且还收集其他学术资料、法院意见和专利。

Google 公司声称拥有除 Elsevier 和美国化学学会以外的所有主要出版商的全文内容，以及 Highwire 和 Ingenta 等托管服务。Google 学术搜索的大部分索引来自商业和开放源代码发布者提供的全文期刊内容的抓取，诸如 OCLC 的 Open WorldCat 和国家医学图书馆的 PubMed 之类的专门书目数据库等。自 2003 年以来，Google 与发布商签订了许多单独的协议，以对全文内容进行索引，而这些内容通常无法通过开放式 Web 访问。

- **百度学术的资源**

百度学术收录了包括知网、维普、万方、Elsevier、Springer、Wiley、NCBI 等的 120 多万个国内外学术站点，索引了超过 12 亿学术资源页面，建设了包括学术期刊、会议论文、学位论文、专利、图书等类型在内的 4 亿多篇学术文献，以

及 300 个学科研究方向，还建设有 400 多万个中国学者主页的中文学者库，覆盖了超过 95%的中国学者；此外，百度学术的期刊库目前已经包含了 1 万多中外文期刊主页的期刊库，300 万科研主题词。

- **AMiner 覆盖的学术资源**

AMiner 学术搜索服务于全球科研人员，覆盖了全球 220 个国家和地区 832 万独立 IP 用户，服务 21 万余家企事业单位及各类机构，提供科研数据下载 230 万次，近 3 年年均数据访问量在 1,100 万次以上。学术资源覆盖全球 1.36 亿学者，超过 2.3 亿篇论文，7.5 亿论文引用关系，879 万知识概念，以及超过 160 个特色专家子库。该平台为中国工程院、国家自然基金委、科技部等科研管理部门提供专家智库、科技发展战略规划等科技情报挖掘服务。

- **Scopus 覆盖的学术资源**

Scopus 收录了来自全球 5,000 多家出版商的 24,000 多种期刊，内容涉及人文、科学、技术及医学等方面，其中同行评审期刊 21,000 多种。Scopus 收录期刊包括来自多个著名的出版商，如 Elsevier、Kluwer、the Institute of Electrical and Electronics Engineers (IEEE)、John Wiley、Springer、Nature、American Chemical Society 等；收录的中国期刊有 578 种，包括《力学学报》、《中国物理快报》、《中华医学杂志》等。此外，Scopus 还收录 800 多种会议录以及数百种丛书。

Scopus 涵盖了 27 个学科领域，归于四大门类：生命科学（4,300 余种）、社会科学与人文艺术（5,300 余种）、自然科学（7,200 余种）和医学（6,800 余种，全面覆盖 Medline）。

通过 Scopus, 用户可以检索到 1823 年以来的近 5,700 万条摘要和题录信息，以及 1996 年以来所引用的参考文献，并且数据每日更新。

Scopus 还提供专利信息和网络学术信息（如预印本、机构仓储、课件库等资源）。

- **Web of Science 覆盖的学术资源**

Web of Science 收录 254 个学科的 20,900 多种高影响力学术期刊，内容涵盖自然科学、工程技术、生物医学、社会科学、艺术与人文等领域。该数据库每周更新。

Web of Science 重要数据库之中，SSCI（社会科学引文索引）和 SCI-E（科学引文索引扩展）的数据均回溯到 1900 年。其中，SSCI 涵盖 58 个社会科学学科的 3,400 多种期刊，以及从 3,500 种世界顶尖期刊中筛选的内容；SCI-E 涵盖 178 个学科的 9300 多种主流期刊。A&HCI（艺术与人文引文索引）涵盖超过 1,800 种艺术与人文领域的期刊，以及从 250 多种自然科学和社会科学期刊中筛选的内容，其数据回溯到 1975 年。此外，CPCI（会议论文引文索引）包括从 180,000 多种会议论文集中获得最前沿、有影响力的研究，数据回溯到 1990 年；BkCI（图书引文索引）截止 2019 年 1 月收录了 101,800 多种图书，同时每年增加 10,000 种新书，数据回溯到 2005 年。

除了上述综合引文索引外，Web of Science 还包括专科引文索引，其中，Current Chemical Reactions 收录了 1985 年以来的最新化学反应，Index Chemicus 收录了 1993 年以来的化学物质的事实型数据，Emerging Sources Citation Index（ESCI）展示了 2015 年以来重要的新兴研究成果。Web of Science 中的这些学科数据库既可以独立使用，也可以综合起来进行检索。

- **中国知网 CNKI 覆盖的学术资源**

中国知网整合了 90%以上的中外文学术文献资源，大规模集成整合传播我国期刊、辑刊、硕士和博士学位论文、工具书、国内外会议论文、报纸、年鉴、国内外专利、科技成果、古籍、图片、文艺、文化、科普、党建、政报公报、经济信息期刊等各类文献资源的大型全文数据库和二次文献数据库。文献量逾 4 亿篇，时间跨度数百年，是实现用户一站式获取学术信息与情报的数据基础。同时实现知识资源的非线性整合传播，全方位满足科学研究、临床应用、经济管理、政策决策等需求，为最广泛的机构、个人用户提供知识管理、学术创新、数字化学习等服务。

总库整合出版中文学术文献资源约 1.2 亿篇(条)，每日更新 1~3 万篇(条)，覆盖 12 种资源类型，最早回溯至 1915 年，基本把有文字记载的学术文献资源

收集完备。同时整合出版外文文献数据库，如 Springer，Elsevier 数据库等，已合作出版 400 余种外文数据库，文献量约 3.0 亿篇（条），外文文献最早回溯至 1836 年，外文图书约 267.6 万种，最早回溯至 1883 年。

4.5 产品代表性研发人才

通过网络公开信息和 AMiner 人才智库，获得 AI 学术搜索领域业界人才情况。本部分选取了领域内代表产品的主要研发或对产品推出具有重要贡献的人才进行展示。

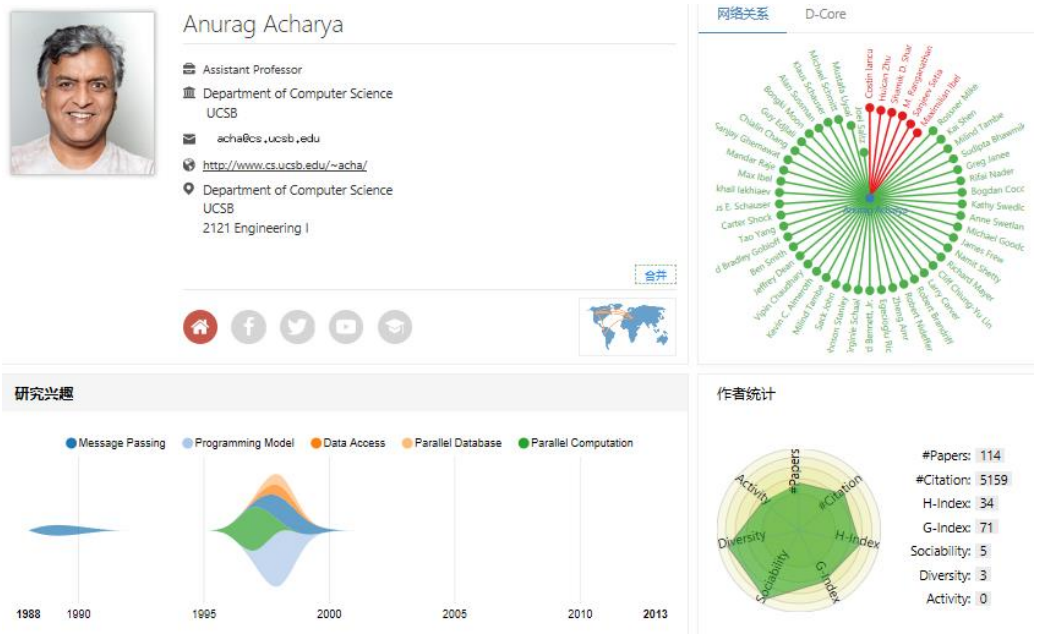
Google Scholar 主要研发人才

Anurag Acharya

Anurag Acharya 是 Google 的杰出工程师，也是 Google Scholar 的创始人之一，曾领导 Google 的索引小组。研究方向是历史数据信息检索、文档评分、数据库、搜索引擎等。

他拥有印度理工学院计算机科学学士学位和卡内基梅隆大学计算机科学博士学位。

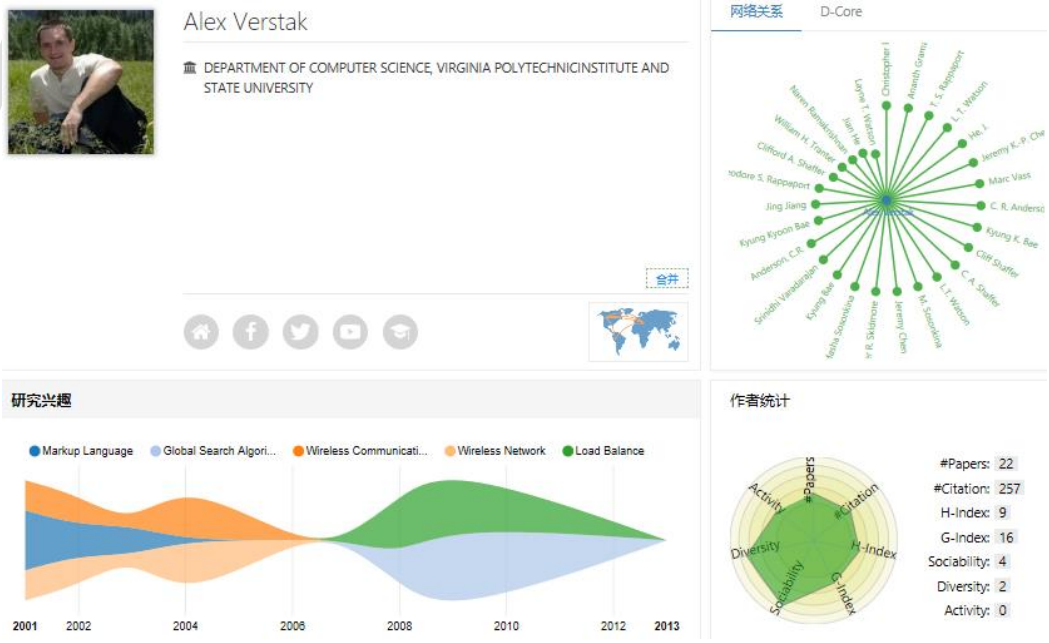
在加入 Google 之前，他是马里兰大学帕克分校的博士后研究员，以及加利福尼亚大学圣塔芭芭拉分校的助理教授。2000 年，Acharya 加入了 Google 索引编制团队他在那里工作了四年。此后，他与 Alex Verstak 一起分析 Google 用户如何作为纯学术咨询来源，并且致力于元数据的自动抽取以及科学文献的定位和排序，以便将结果整合到通用搜索引擎中，提高网络搜索中学术文件的排名。



Alex Verstak

现任谷歌工程师，参与创建了 Google Scholar。研究方向是计算机网络、全局搜索算法、全局优化、分层数据挖掘等。

2002 年毕业于弗吉尼亚理工大学获得计算机科学硕士。毕业后即加入谷歌。



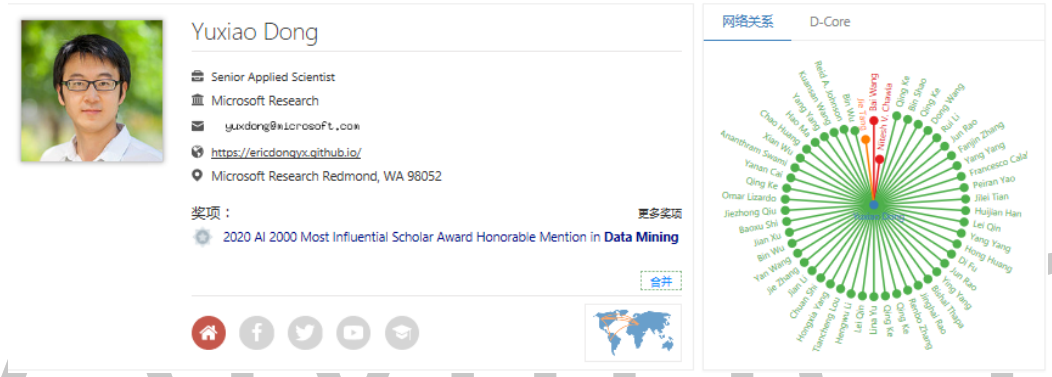
微软学术主要研发人才

Yuxiao Dong

现任 Redmond 微软研究院高级应用科学家。研究方向是社交和信息网络，数据挖掘和应用机器学习，重点是将计算模型应用于解决大规模网络系统中的问题，例如 Microsoft Academic Graph（MAG）、在线社交媒体和移动通信。

2017 年获得了圣母大学的计算机科学博士学位。

2011 年获得 ADMA'11（第七届高级数据挖掘和应用国际会议）最佳应用论文奖，2013 年获得 DKE（数据和知识工程）高引用研究奖，2015 年获得 WSDM'15 最佳论文奖提名。



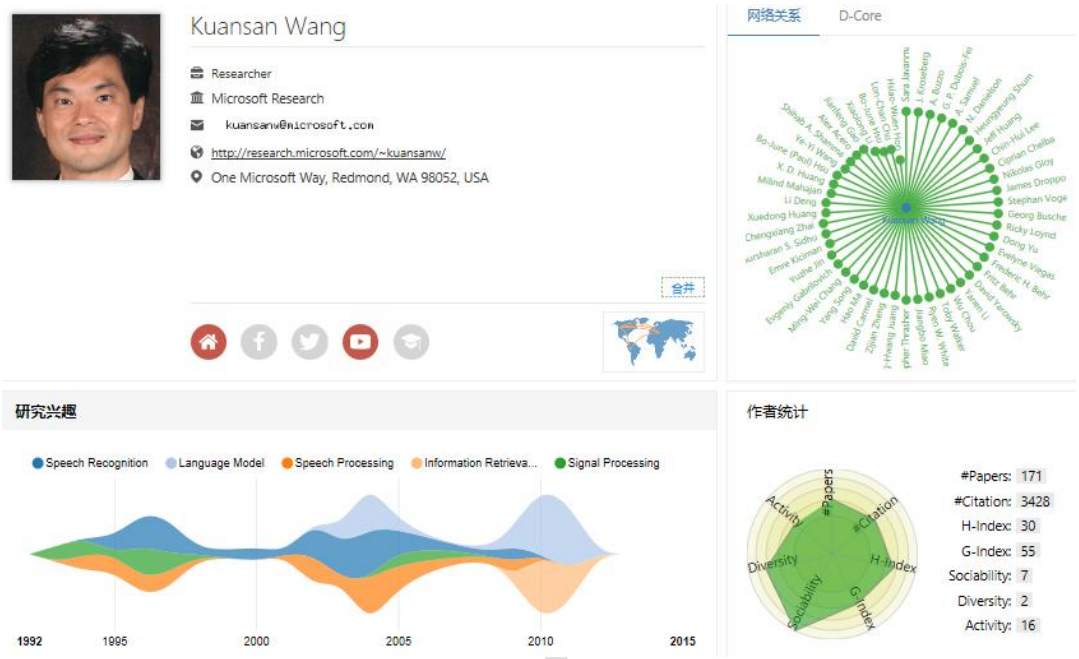
Kuansan Wang

现任 Microsoft Academic Services 总经理

1994 年获得马里兰大学电机工程博士学位。

Kuansan 于 1998 年 3 月加入微软研究院，曾任语音技术小组研究员，从事口语理解和对话建模领域的研究；2004 年 1 月转到语音产品组，成为软件架构师帮助创建并发布了产品 Microsoft Speech Server，还参与了万维网联盟（W3C）语音识别语法规范（SRGS）、W3C 语音合成标记语言（SSML）以及 W3C 多模式交互工作组的各种其他出版物；2007 年 9 月加入了新成立的互联网服务研究中心，2010 年加入必应项目，2016 年 3 月担任 Microsoft Academic Services（MSR）外展部的总经理，还创建了一个实验网站 academic.microsoft.com（由学术 API 提供支持）和移动应用程序，同时他作为创始成员之一推出了微软 Response Point，一个支持语音的小型企业电话系统。

在加入微软之前，曾在贝尔实验室（Bell Labs）和纽约大学（NYNEX）的科学技术中心工作。

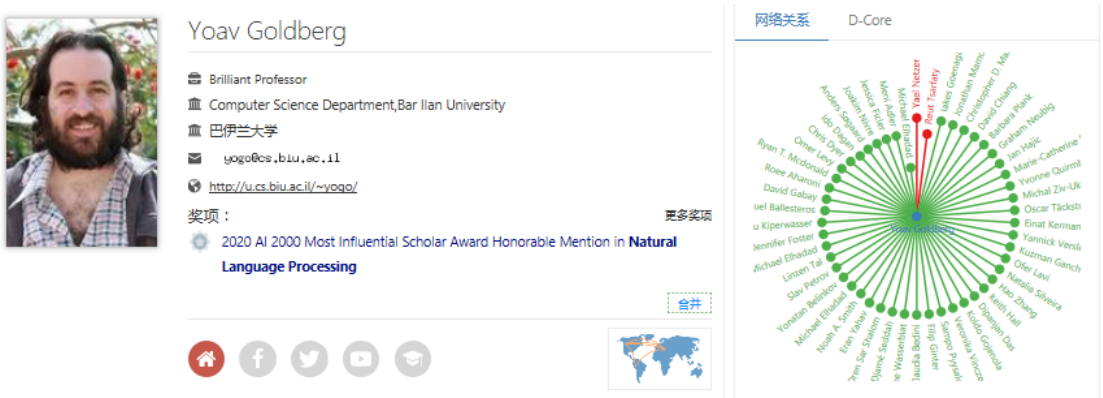


AI2（Allen 人工智能研究院）语义学术（Semantic Scholar）主要研发人才

Yoav Goldberg

现任 AI2 以色列的研究主任，也是巴伊兰大学（Bar Ilan University）计算机科学副教授。他的研究兴趣包括语言理解技术、符号和神经表征语言、揭示文本、句法和语义处理的潜在信息，以及解释性和基础理解的文本和序列深入学习模型。

曾撰写了一本关于自然语言处理深度学习技术的教科书，2018 年跻身 IEEE 人工智能十大观察对象，并在 2017 年获得 Krill Prize 科学奖。

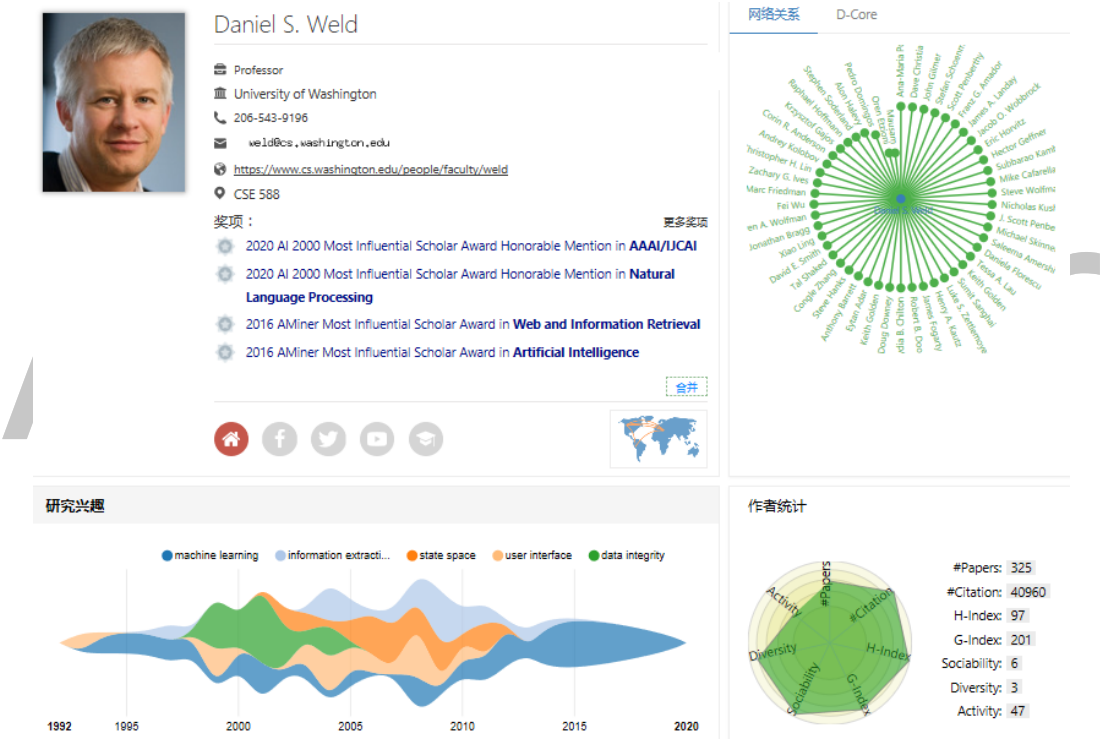


Dan Weld

现任 AI2 的 Semantic Scholar 研究团队管理者，同时也是华盛顿大学 Paul G.Allen 计算机科学与工程学院的 Thomas J.Cable/WRF 教授。

1982 年获得耶鲁大学获得学士学位，1988 年获得麻省理工学院博士学位。

Weld 是人工智能进步协会（AAAI）和计算机械协会（ACM）的成员，获得了总统青年调查员奖、海军研究办公室青年调查员奖和多项最佳论文奖。Weld 与他人共同创立了几家公司，包括 Netbot Incorporated（由 Excite 收购）、AdRelevance（由 Nielsen NetRatings 收购）和 Nimble Technology（由 Actuate 收购）。

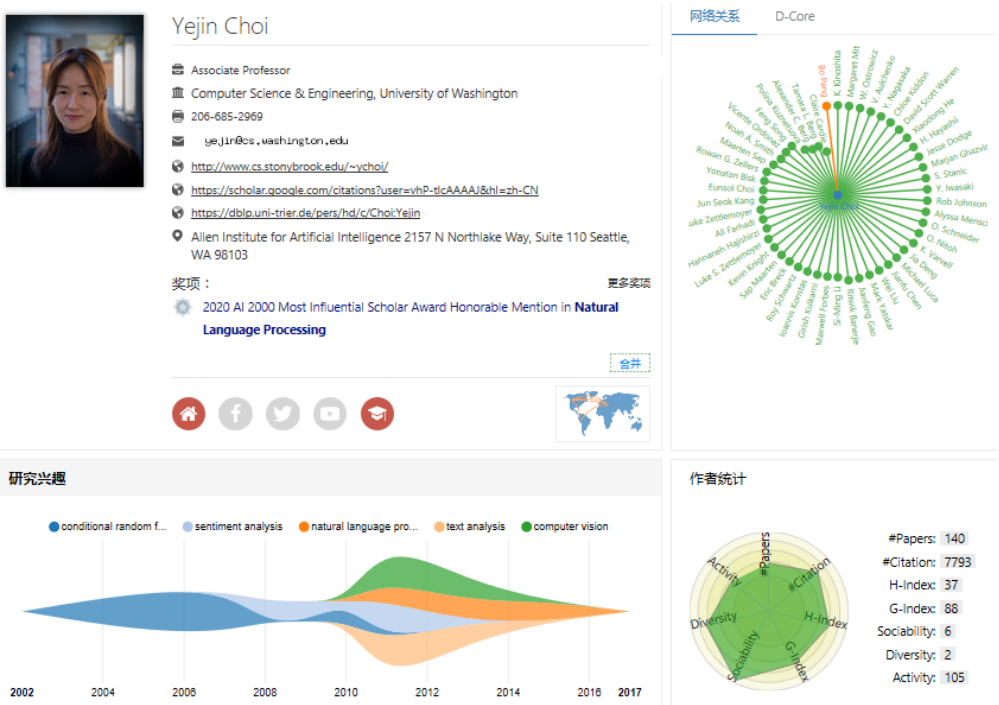


Yejin Choi

现任美国艾伦人工智能研究所的高级研究经理，也是华盛顿大学 Paul G. Allen 计算机科学与工程学院的副教授。研究领域：自然语言处理、机器学习、人工智能以及计算机视觉和数字人文科学。

拥有韩国首尔国立大学计算机科学与工程学士学位，康奈尔大学计算机科学专业博士学位，师从 Claire Cardie 教授。

2013 年获得 ICCV 的 Marr 奖（最佳论文奖），2016 年被 IEEE AI 评选为十大值得关注的人，2018 年获得博格早期职业奖（BECA）。

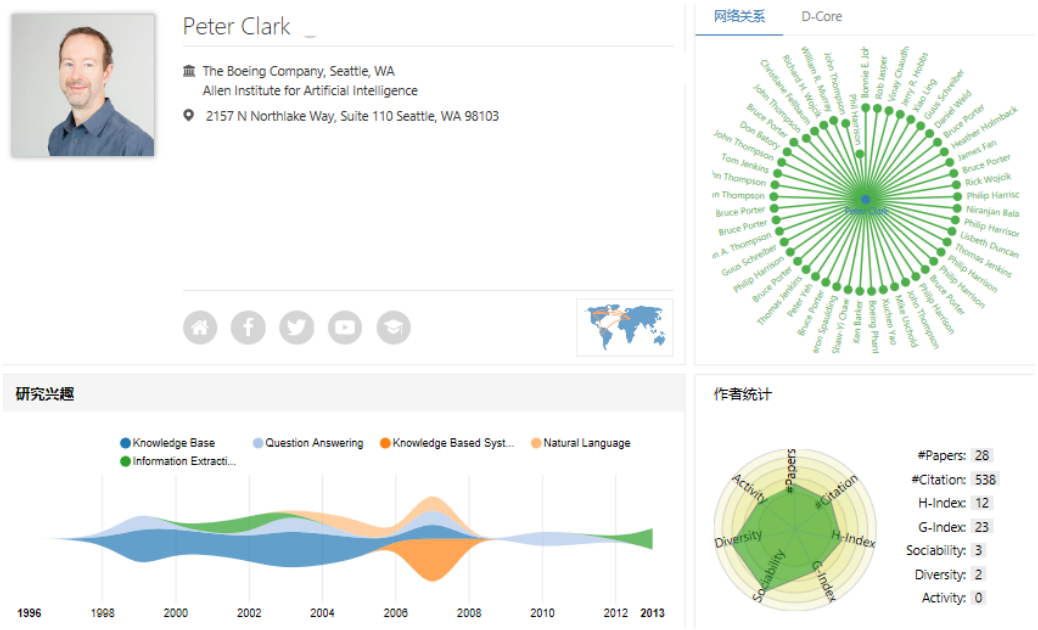


Peter Clark

现任美国艾伦人工智能研究所的高级研究经理。他的工作重点是自然语言处理，机器推理和大型知识库，以及这三个领域之间的相互作用。研究领域是自然语言理解、学术数据库、机器推理和常识推理。

1991 年获得计算机科学博士学位。

获得奖项包括 AAAI 最佳论文（1997）、波音公司技术研究员（2004）和 AAAI 高级会员（2014）。至今已研究自然语言处理、数据库领域 30 余年，发表 80 余篇出版物。



AMiner 主要研发人才

唐杰

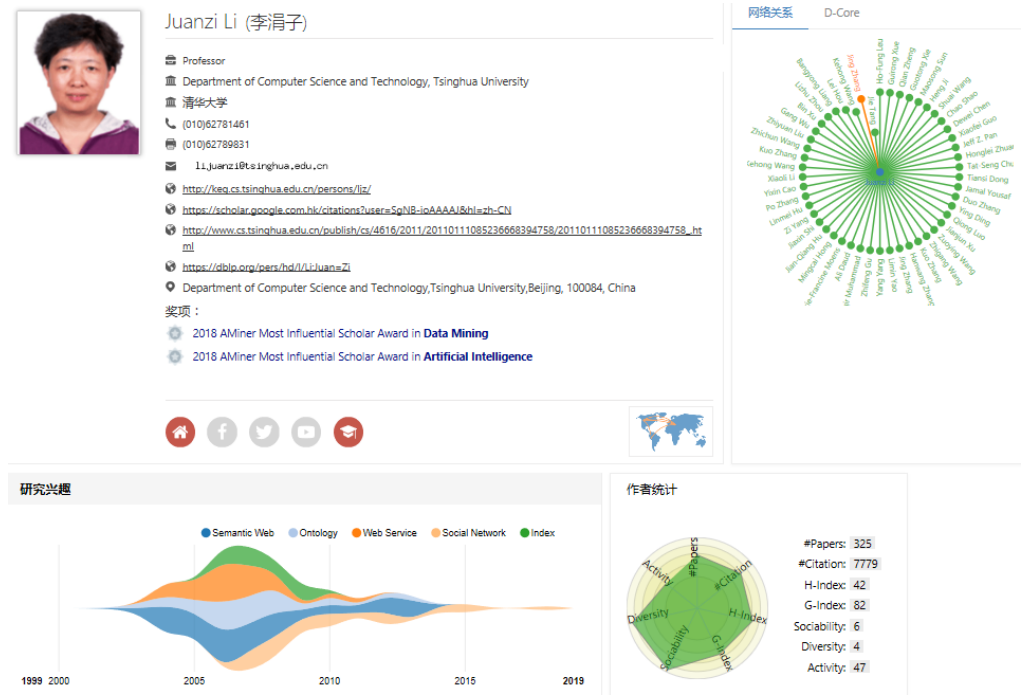
清华大学计算机科学与技术系长聘教授，计算机系副主任、清华-工程院知识智能联合实验室主任。研究兴趣包括：社会网络分析、数据挖掘、机器学习和知识图谱。

发表论文 200 余篇，拥有专利 20 余项。主持研发了科技情报大数据挖掘与服务系统平台 AMiner，吸引了 220 个国家/地区 1,000 多万独立 IP 访问。曾担任国际期刊 ACM TKDD 的执行主编和国际会议 CIKM'16、WSDM'15 的程序委员会主席、KDD'18 大会副主席以及 IEEE TKDE、ACM TIST、IEEE TBD 等期刊编委。获英国皇家学会-牛顿高级奖学金、CCF 青年科学家奖、国家自然科学基金委员会杰出青年学者、北京市科技进步一等奖、中国人工智能学会科技进步一等奖、KDD'18 杰出贡献奖。2012 年获得国家优秀青年科学基金奖。

李涓子

清华大学长聘教授，清华大学人工智能研究院知识智能研究中心主任，中国中文信息学会知识与语言计算专业委员会主任。

在知识工程研究领域取得了突出成果，主持研发了基于语义链接的跨语言知识图谱 XLORE，参与研发了科技情报大数据挖掘与服务系统平台 AMiner，曾获北京市科技进步一等奖、人工智能学会科技创新一等奖、王选新闻科学技术进步一等奖等多个奖项。



BASE 主要研发人才

Dirk Pieper

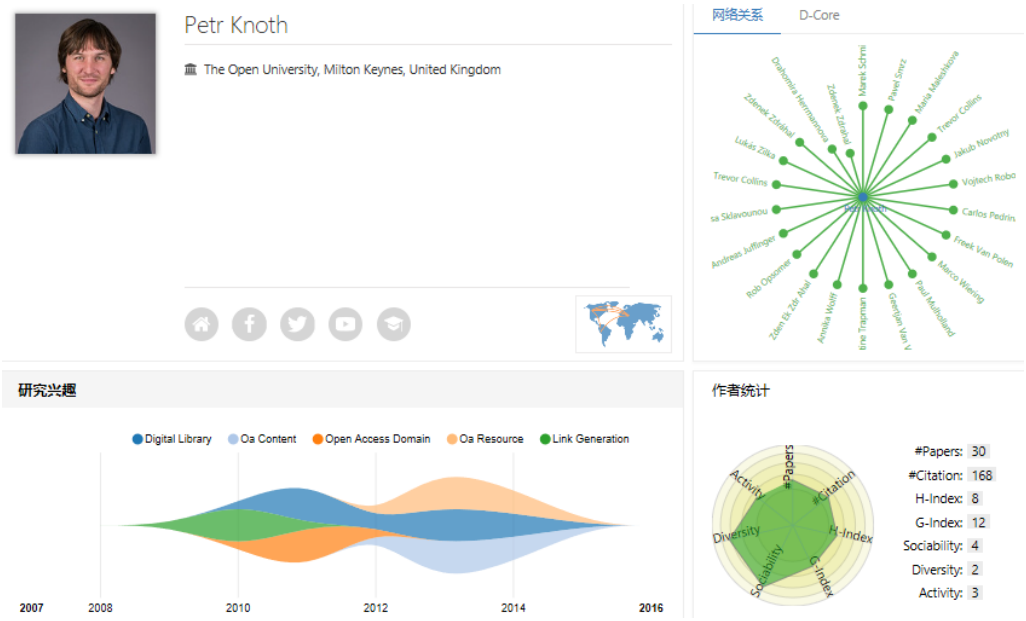
BASE 项目负责人，Bielefeld 大学图书馆总办公室副主任。

CORE 主要研发人才

Petr Knoth

CORE 创始人兼负责人，负责领导团队进行研发、资金和产品策略。研究方向：自然语言处理、文本和数据挖掘开放访问、开放科学、学术交流信息检索、信息抽取、推荐系统、科学计量学等。

曾在文本挖掘、开放式科学和电子学习等领域的许多欧盟委员会、国家和国际资助的研究项目中担任研究员。曾在 Mendeley 担任高级数据科学家，负责信息抽取和研究内容推荐。他对使用 AI 改善研究工作流程非常感兴趣，与他人共同创立了 Semantometrics.org，其目标是超越文献计量学和高度测量学，以产生新的研究评估方法。他在 2010 年开发了第一个 CORE 原型。



Web of Science 主要研发人才

Keith Collier

现任产品管理副总裁。

Collier 拥有超过 23 年的出版行业经验，此前，Collier 曾在汤森路透旗下的 ScholarOne 公司担任负责人，该公司为学术期刊、书籍和会议提供全面的工作流管理系统。

知网主要研发人才

王明亮

教授级高工。同方股份副总裁、同方知网董事、总经理，中国学术期刊（光盘版）电子杂志社董事长。

1989 年清华大学物理系硕士毕业。

曾任山西大学理论物理研究所讲师，在国内外发表论文 10 余篇。1995 年与清华物理系合作创办北京清华信息系统工程公司，是中国知网主要创始人。获“新中国 60 年百名优秀出版人物”等称号。

MingLiang Wang

董事长、高级副总裁

同方股份高级副总裁、中国知网董事长

Beijing,China

4.6 学术评价指标

随着文献数据量的不断扩大，以及数据库的检索和排序功能日益完善，对期刊、文献或学者进行学术影响力评价的需求越来越大，也越来越受到学术界的重视。自 20 世纪 90 年代以来，学术搜索产品逐渐开始利用自身不同数据源的文献及引文数据等构建出了各种评价指标模型。

评价分析功能是指学术搜索引擎或数据库利用各种引文数据对论文、期刊或作者等进行学术影响力评价的功能。这些学术评价指标**主要分为期刊评价指标、文献评价指标和作者评价指标三类。**

期刊评价指标主要评价该期刊自创刊以来所登载的全部论文在统计当年被引用的情况，用于显示该期刊被使用和受重视的程度，以及在科学交流中的作用和地位。目前，较为知名的期刊指标评价有 Web of Science 中的期刊影响因子、Scopus 的 CiteScore 和 Google Scholar 的谷歌学术计量等。

文献评价指标主要用于对某篇文献的质量和学术成果影响力进行评价。目前，较为知名的文献评价指标有 Google Scholar 的 GSC、语义学术 Semantic Scholar 的文献影响力评价和百度学术指数。

作者评价指标主要用于该科研作者的学术实力和影响力，较为知名的作者评价指标有作者 h 指数、AMiner 学者指标分析等。

4.6.1 学术期刊指标评价

Web of Science 的期刊影响因子评价指标

影响因子是汤森路透出品的**期刊引证报告**（Journal Citation Reports，简称 JCR）中的一项数据，已成为评价论文质量，评价期刊、科研机构、科研人员学术水平的重要指标。

JCR 是一个多学科期刊评价工具。统计 Web of Science 核心合集收录期刊所刊载论文的数量、论文参考文献的数量、论文的被引用次数等原始数据，再应用文献计量学的原理，计算出各种期刊的影响因子、立即影响指数、被引半衰期等反映期刊质量和影响的定量指标。

自 1975 年以来，每年 6 月份公布。是以 SCI 期刊为对象，统计在一定时期（通常是前两年）内，某一刊物发表的论文被已经进入 SCI 刊物的论文所引用的总次数，除以该刊物这一时期内的论文总数，即，期刊在过去两年发表的论文在当前 JCR 年的平均被引次数。

$$\textbf{IF（报告年份）} = \frac{\text{该期刊前两年发表的论文在该报告年份中被引用总次数}}{\text{该期刊在这前两年内发表的论文总数}}$$

CiteScore

Scopus 的**期刊和丛书的 CiteScore 度量标准，**以三年区间为基准来计算每本期刊的平均被引用次数。

谷歌学术计量（GSM）

Google Scholar 于 2012 年推出**谷歌学术计量**（Google Scholar Metrics,简称 GSM），用来评价各个领域杂志的影响力。从 2012 年起来，GSM 每年都会发布学术期刊和会议的 GSM 排名。

GSM 主要包括 **h 指数（h-index）、h 核心（h-core）、h 中值（h-median）**三个指标。

- h 指数：指在所有发表的论文中，有至少 h 篇论文分别被引用了至少 h 次，那么这份期刊或会议的 h 指数就是 h。
- h 核心：指该期刊或会议被引用最高的 h 篇论文。

- h 中值：指 h 核心中位数论文的被引用次数。

GSM 排名方法主要采用 **h5 指数（h5-index）**、**h5 核心（h5-core）**和 **h5 中值（h5-median）**。**h5 指数可以体现期刊和会议的整体综合实力，逐渐成为学术出版物和会议影响力评价的一个重要参考**。

h5 指数是指通过计算过去五年中出版物的 h 指数来说明学术出版物中最新文章的知名度和影响力。相应地，h5 核心和 h5 中值，是指收录在谷歌学术系统中的期刊和会议在最近五年被引用最高的论文数量及中位数论文被引用的次数。

4.6.2 论文评价

Google Scholar Citations（简称 GSC）

GSC 为作者提供了一种简单的方式来跟踪自己文章的引用情况。

（1）作者使用 Google 个人帐户自动创建和编辑个人资料；

（2）作者可以查看是谁引用了自己的出版物，以图表形式查看各个时段的引用情况，并计算多项引用指标；

（3）作者可以将个人学术档案公开，这样，其他人搜索自己的姓名时，自己个人学术档案就可以显示在 Google 学术搜索的结果中；

（4）导出书目记录，用于构建作者个人主页。

Semantic Scholar 的文献影响力评价指标

语义学术 Semantic Scholar 提出**一套新的文献影响力评价指标**，根据该文章对施引论文的重要性，而非仅根据被引量和下载量两个指标。Semantic Scholar 使用深度学习技术为学术文献设计了如高影响力引用、引用加速度、年度引用趋势图等文献、作者评价指标，并运用深度学习技术优化推荐排序，旨在用尽可能少的检索次数检索到更符合其信息需求的学术文献。其中，高影响力引用指标主要用于识别被引文献对引用文献的重大影响，它是基于机器学习模型来分析一个文献的引用次数以及引用语境等多个因素而确定的，可以帮助人们更容易理解文献之间是如何建立联系的。

百度学术指数

百度学术指数包含根据学术成果计算的基础指数（总被引频次、总成果数、h 指数、g 指数）及被其他人关注的次数即搜索指数。

4.6.3 学者评价

h 指数

h 指数是 2005 年由美国加利福尼亚大学圣地亚哥分校的物理学家乔治·希尔施（Jorge E. Hirsch）提出的。

h 指数是一种评价科研人员学术成就的方法。一个科研人员的 h 指数为在一定期间内他发表的论文至少有 h 篇的被引频次不低于 h 次。

h 指数也可以用来衡量一个机构的学术成就。一个机构的 h 指数指一定时期内该机构的论文至少有 h 篇的被引频次不低于 h 次。该指数计算方法：一个机构的论文数量包括（a）由当前属于该机构的学者所发表的论文；（b）该论文发表时相关作者属于该机构。

g 指数

g 指数是基于科研人员被引次数的分布来评价科研人员学术成就的另一种方法。该指数是 h 指数的衍生指数，由 Leo Egghe 于 2006 年提出，主要是弥补 h 指数不能很好反映高被引论文的缺陷的。

一个科研人员的 g 指数指他的 g 篇被引次数最多的论文平均有 g 次被引，g 是可能的最大数目。

和 h 值一样，g 值越大说明该学者的学术影响力越大、学术成就越高，通常作为 h 指数的补充或提高。

AMiner 学者指标分析

AMiner 产品的学者学术能力表达为多维学术指标描述的雷达图（如图 20）直观展现。学术指标包括：论文数、论文引用数、h 指数、g 指数、学术活跃度指数、学术社交性指数、研究多样性指数等。

学术活跃度指数，是指该作者在一定时期内发表论文的数量排名指数。

学术社交性指数，是指该作者在一定时期内与其他学者合作发表论文的排名指数。

研究多样性指数，是指该作者在一定时期内、在不同研究话题上发表论文的指数排名。

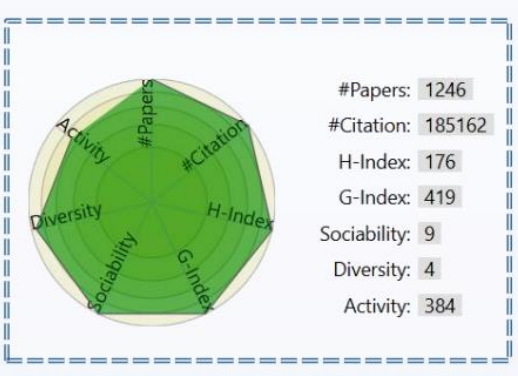


图 20 AMiner 学者指标雷达图

4.7 产品功能和技术

本部分主要从产品功能角度对当前主要的学术搜索产品进行分类展示，并分析了人工智能技术在相应产品中的应用。

4.7.1 多源异构数据融合与命名排歧

多源异构数据具有类型混合、离散、数量大、质量不齐等特点。首先，多源异构数据通常是既包括结构化又包括非结构化数据的混合型数据；其次，数据因分布在不同的系统或者平台而具有离散性特点；还有，各个平台所包含的数据量都非常庞大，但是不同平台上的数据质量则是参差不齐。

在学术搜索领域，对多源异构数据进行融合，并基于融合后的数据提供文献搜索服务，是许多产品共同追求的目标。

目前，几乎所有的学术搜索产品都具备多源异构数据融合功能。知识图谱技术使得多源异构数据融合可以通过算法实现。学术文献中通常蕴含了丰富的信息，如研究主题、发表期刊、论文引用情况、作者、合作者关系网络、作者在某时段的工作单位或研究机构、受基金项目资助情况、刊物的覆盖内容和学术会议召开

的时间及地点等。对这些信息进行深入分析，可构建出关系丰富的知识图谱。从不同数据源构建图谱的时候，利用自动化的算法可以将各处的搜索实体进行聚类。

多源异构数据融合过程中，存在一个命名消歧的问题。处理作者名歧义问题主要有两个困难：一是如何量化来自不同数据源实体间的相似性；二是如何确定具有相同姓名的作者人数。目前，**仅有微软学术、Scopus、AMiner 等少数产品能够实现了命名消歧功能。**例如，Scopus 通过算法来识别同一作者在不同时期、不同研究机构发布的文章，从而进行**作者身份识别。**

4.7.2 一般检索与高级检索

学术搜索产品基本功能相互类似，但是，在 AI 技术引入程度方面存在差异。**除了 Semantic Scholar 因 AI 特色独特而仅提供单一检索选项之外，其他学术搜索产品均都同时提供一般搜索、高级搜索两个选项。**

一般搜索通常可以是关键词、作者、期刊图书名称、期刊号等进行，几乎所有的学术搜索产品都支持该功能。例如，Web of Science 的简易检索共有三个检索入口，即：主题、人名、地名检索。

关于高级搜索，各个产品给出的搜索指定域通常有一些差异。例如，Google Scholar 的高级搜索提供可搜索文章或书籍标题、作者姓名、期刊标题、出版年份，或者根据关键字检索文档或页面匹配项提供全文检索。微软 Academic Search 和 CORE 的高级搜索可以限制作者、会议、期刊、组织、年代、学科和 DOI（Digital Object Identifier，数字对象唯一标识符）。AMiner 的高级搜索允许根据关键词、作者、机构等进行。ScienceDirect 的高级搜索提供在指定域的搜索，如 Search in author keywords（作者关键词）、Search in text only（正文检索）、作者所在机构、期刊号、页码、参考文献，学科分类、文章类型、语种等。

百度学术 Baidu Xueshu 的高级检索支持按关键词/主题、标题、DOI 等进行检索，还支持用户设定包含/不包含精确检索词、出现检索词的位置、作者、出版物、发表时间、语言检索范围，也可以利用高级语法直接进行检索。此外，百度学术还支持参考文献串检索，即当用户的输入词为参考文献格式表示的一串内容时，搜索结果能够自动分析该格式，找到用户寻找的目标文献。

BASE 的高级搜索支持选项更加多样化。用户不仅可以指定按文献全文、题名、作者、ORCID ID、关键词组、DOI、部分 URL、搜中页面、国家来源、出版年份等不同条件搜索，而且可以指定是否优先考虑开放获取文献、条款/许可证，以及指定文本、图书（含章节）、期刊报纸、会议、评论文章、手稿、专利、学位论文、乐谱、地图、图片、音视频、软件、数据等不同文献类型，还可以指定精确搜索、相近词义、或多语种搜索等语言工具。

Science.gov 的高级搜索支持全文搜索，以及按标题、作者、年份等进行，并且，**可按大类或主题词检索**，可以同时检索该网站的数据库与网址。大类包括：农业与食品、应用科学与技术、航天与宇宙、生物与自然、保健与医学、能源、计算机与通信、环境、地球与海洋、数理化、自然资源、科学教育等 16 个。

Web of Science 提供综合检索与引文检索两大途径。综合检索是对来源文献的检索，包括主题检索、著者检索、来源期刊检索和著者地址检索（著者地址检索可按机构名、城市名、国名甚至邮政编码检索）四大项，结果显示控制在 500 条之内；引文检索的入口包括被引文著者、被引文献（如被引期刊缩写、被引书名或被引专利号等，但不得超过 20 个字符）和被引文年代三项。

中国知网的高级检索支持按文献分类目录、主题、关键词、篇名、摘要、全文、被引文献、中图分类号、DOI、发表时间、文献来源、支持基金等进行搜索；支持模糊匹配、词频限定等。其次，知网还提供**专业检索**，用户可以使用逻辑运算符合关键词构造检索式进行搜索，用于图书情报专业人员查新、信息分析等工作。支持**作者发文检索**，通过作者姓名、单位等信息，查找作者发表的全部文献及被引下载情况。知网还支持查找同时包含两个关键词的**句子检索**和**一框式检索**。

不同于其他学术搜索产品，Semantic Scholar 依赖于其强大的 AI 技术而仅提供单一检索选项。Semantic Scholar **利用“机器阅读”技术从文本中挑选出最重要的关键词和短语**，而且不需要依赖作者或出版商键入这些关键词。此外，它可以**判断文章所论述的主题**，也可以**从论文中抽取图表**，将它们呈现在检索结果中，帮助用户快速理解论文内容。**辨别一篇文章引用的参考文献是否具有重要的参考价值**，基于此评价论文的学术影响力，可快速获得重要文献。

Semantic Scholar 用户也可以进行**过滤筛选，**通过筛选最近 5 年、文献、是否有 PDF、是否有视频、出版物类型、作者、期刊会议名称等选项更加精准地找到所需文献。

百度学术 Baidu Xueshu 文献检索支持**标题检索，题录区**包含文献的作者、出版源（时间、期卷等）、被引量信息，均支持进一步点击，查看对作者、出版源的二次检索结果及该文献的引证文献。此外，支持**期刊检索**，用户可按期刊名称或在期刊库中检索，结果显示该期刊的影响因子、发文量、出版周期、搜索指数、被引量等，并链接到官方网站，用户可在线投稿、交流讨论。百度学术还支持**中英互译：**当用户输入中文关键词时，提供中英互译功能支持用户一键搜索相应的英文文献，无需翻译和二次输入。

4.7.3 搜索结果显示

目前的学术搜索产品在结果显示、全文获取方式、文献引用和导出等方面的功能多数类似，仅存在一些细微差异。如表 5 所示。

表 5 学术搜索产品主要功能对照表

	Google Scholar	Microsoft Academic	BASE	CORE	Science.gov	Semantic Scholar	Baidu Scholar	AMiner
摘要	部分显示	✔	✔	✔	✔	✔	部分显示	✔
相似文章	✔	✔	✘	✔	✘	✔	✔	✔
参考文献	✔	✔	✘	✘	✘	✔	✔	✔
被引用数	✔	✔	✘	✘	✘	✔	✘	✔
全文链接	✔	✔	✔	✔	✔	✔	✔	✔
导出格式	APA, MLA, Chicago, Harvard, Vancouver, RIS, BibTeX	APA, MLA, BibTeX	RIS, BibTeX	BibTeX	APA, MLA, RIS, BibTeX	APA, MLA, Chicago, BibTeX	APA, MLA, RIS, BibTeX	APA, MLA, Chicago, BibTeX

Google Scholar 的检索结果页面还包含如下的深入检索链接。

✓ “**Cited By**”：搜索引用该文的其他论文。

- ✓ “**Library Links**”：链接到用户所属图书馆的资源；主要为学生或科研人员等机构用户提供附加价值，因为谷歌使用开放链接/链接解析程序（如SFX）直接链接到当地图书馆馆藏。
- ✓ “**Library Search**”（库搜索）：搜索拥有该图书的图书馆；或将查询链接到 OCLC WorldCat，提供本地库的点击率，还将呈现文档在 web 上的替代位置。
- ✓ “**Alternate Version**”：搜索该论文的其他版本，如预印本、摘要、会议论文等。
- ✓ “**Web Search**”：通过 Google 搜索有关此论文的其他信息。特别是当文档不能直接从 Google Scholar 结果列表中获得，并且查询扩展到整个（Google）网络时。

微软/必应学术的搜索结果页面包括论文题目、引用量、作者姓名、论文摘要、出版时间和论文来源等。其中，论文题目可点入链接到**论文详细信息页面**，显示该论文的作者信息、摘要、参考书目、引用书目等；作者姓名可点入链接到作者详细信息页面，显示作者的个人主页、所属机构、论文列表、学术引用图等；论文来源可以点入链接到**会议、期刊的详细信息页面**，显示其论文数量、引用数量、论文列表等。

Science.gov 搜索结果页面包括文章列表，包含标题、星级、日期、外部链接、作者、DOI、摘要等。

BASE、CORE 和 ScienceDirect 的搜索结果页面显示文章列表：包含标题、作者、摘要、文献类型、出版年、数据提供者或所在数据库。除此之外，Scopus 检索结果页面还显示文献的被引用次数、以及文章的参考文献信息。

此外，Web of Science 检索结果页面还提供了**参考文献**（Cited References）**与被引文献**（Times Cited）。点击参考文献链接，可得到来源文献的著者撰写论文时所列的所有参考文献的著录；显示文章的来源文献被引用次数，点击链接可得到该文章被别人引用的相关文献的著录。若其相关文献同时是 SCI 来源文献库中的记录，则可检索到该文献的详细记录。

语义学术 Semantic Scholar 搜索结果页面**结果列表**包含文章列表，作者链接、学科、发表年份、高影响引用值、引用速度、出版商该文链接，导出引用、保存到个人图书馆、相似主题等。**文章页面**包含该文章的标题、作者、摘要、出版商、链接到图书馆、所属学科、发表年份、DOI、被引用情况（背景、方法、结果），并支持导出引用。**作者页面**显示论文发表量、引用量、高影响力引用量，以及所有论文。该作者论文可按年份、合作者、期刊会议名称等进一步筛选。

4.7.3.1 搜索结果排序

对于多数学术搜索产品，其搜索结果显示通常包括文献标题、摘要、被引用数、作者等列表信息，并且支持按照相关性、时间、作者、标题等规则进行组织排序显示。排序考虑到了每一篇文章的全文以及文章的作者、文章出现的出版物以及在学术文献中被引用的频率等因素。最相关的结果将显示在第一页、最有用的参考出现在页面顶部。

其中，Web of Science 检索结果支持多种排序。

- ✓ 按 ISI 收到文献并处理的日期降序排列，时间越近，排序越前。检索结果最多提供 500 条。
- ✓ 按被引次数降序排列，被引次数越多，排序越前。检索结果最多提供 300 条。
- ✓ 按相关性排序，即检索词匹配的频率越高，排序越前。检索结果最多提供 500 条。
- ✓ 按第一著者名升序排列，匿名著者排在最前面。检索结果最多提供 300 条。
- ✓ 按来源期刊刊名升序排列。检索结果最多提供 500 条。

百度学术**排序功能**支持按相关性、被引量、时间降序三种方式将文献进行排序，默认为按相关性排序。**筛选功能**支持按发表时间、所属研究领域、核心数据库收录情况、包含关键词、文献类型、作者、发表期刊、发表机构八种方式将文章进行细粒度的筛选，缩小搜索范围，找到所需文献。

4.7.3.2 全文获取方式

学术搜索产品致力于为用户提供公开免费的文献全文。对于一些版权文献，搜索产品通常提供文献摘要以及相应的付费入口链接。Google Scholar **免费显示文档摘要（而非全文），并附带按次付费观看选项。**Google Scholar 索引的大部分内容都存储在出版商的服务器上，在那里可以免费下载全文文档。

微软/必应学术通过**下载论文**链接到论文浏览/下载页面。Scopus 通过 SFX 链接获取全文。CORE 的文献提供 PDF 获取。BASE 的文献可以进行**浏览**，用户可按杜威十进分类（DDC）、按字母或标号顺序的文献类型、使用协议条款、开放/不开放获取方式进入 4 种方式进行浏览。

百度学术通过**下载区**给出搜索结果文献的全部来源供用户选择进行下载，用户也可从“免费下载”tab 下直接查看所有的免费来源进行选择，若无法获取到该文献可以选择向其他用户求助。同时，百度学术**文献下载**收录了同一篇文章的多个来源，并在用户检索步骤中提供在检索结果列表中、文献详情页等多个下载入口。

4.7.3.3 相似文章推荐

相似文章推荐功能在学术搜索产品之中较为常见。AMiner、ScienceDirect 等产品直接在搜索结果页面**推荐相似文章。**

Web of Science 通过**相关记录框**（Related Records）进行推荐。通过引文同被引的线索，把有关来源文献联系起来。通过点击相关记录键，可以找到在不同年份中共同引用某些参考文献的相关文献。相关记录按照引文同被引的数量降序排列。

百度学术则专门设置了**推荐区**，支持用户按照相似文献、参考文献、引证文献三个维度查看更多相关文献。百度学术会综合考虑文献的相关性、权威度、时效性等多维度指标，提供与输入词最相关的多篇文献。

4.7.4 专家检索与审稿人推荐

专家检索是指利用能够表征专家专长的各种文档和资源，识别专家在某给定查询主题（领域）的专长（相关性）程度，并按程度高低排序显示专家结果列表的过程。

微软学术/必应学术的**学术实体详细页面**：包括作者、机构等。通过**作者姓名**链接到**作者详细信息页面**，显示作者的个人主页、所属机构、论文列表、学术引用图等。

Web of Science **机构数据中介页**（intermediary page）：用于进行数据中介（mediate access）以及跟踪其所服务的各分支机构的使用状况。

百度学术提供**学者检索/学者库**，相应的学者检索结果页面包括：

- ✓ **ScholarID**：学者主页的唯一标识码。
- ✓ **全部学术成果**：所有已发表或未发表的学术成果，支持按年份、文献类型、是否第一作者进行筛选，支持按发表时间和被引量进行排序。
- ✓ **学术成果分析**：将全部学术成果从多个维度进行分析，包括成果类型分析、年度成果数分布、年度被引量分布。
- ✓ **学术指数**：包含根据学术成果计算的基础指数（总被引频次、总成果数、h 指数、g 指数）及被其他人关注的次数即搜索指数。
- ✓ **合作分析**：运用大数据技术对学者的学术成果进行分析，用户可方便查看到高频的合作学者及合作机构。

AMiner 学术搜索也具有专家检索功能，同时还具有**审稿人推荐功能。**该产品以数据为支持、采用知识图谱技术分析、用人工智能技术自动生成全球 AI 最有影响力学者榜单、全球计算机领域高校排名、全球学术会议综合指数及排名等**学术榜单**。此外，AMiner 还支持**精准推送**功能，通过高质量语义分析和内容生成技术，结合专家画像，在 AMiner 平台覆盖的全球 1.36 亿学者中帮助期刊在交叉领域和新兴热点领域寻找国内外合适的约稿对象和审稿人。

4.7.5 网络关系分析

随之社交网络应用的普及，学术网络关系分析功能需求越来越多。相应的，学术搜索产品陆续推出了网络关系分析服务。

AMiner **学术专家画像**系统，按照院系、研究兴趣、影响力、奖项、区域等多重维度将人才进行分类管理，建立了**智能人才数据库**。并且，通过科研人员公开发表的论文、专著、专利等信息生成了**学者网络关系图**，为高校提供更加清晰地科研人员关系网，为发掘科研工作者的师生关系、团队关系、合作关系等提供新的方法和手段。如图 21 所示。

图 21 AMiner **学术专家网络关系**

该搜索产品的网络关系分析还包括了学者的**代表性成果展示。**科研人员的代表性成果由学者的代表性论文、牛人引用、高被引论文和成果影响力组成，帮助科研管理人员挖掘科研人员的学术亮点。挖掘并展示了科研人员所有有影响力的科研**奖项与荣誉**，以供参考。如图 22 所示。

图 22 学者代表性成果与荣誉奖项展示

此外，网络关系分析还支持展示**专利与基金项目**（如图 23 所示），例如，国家自然科学基金和人才培养项目是国家科技体系中的重要组成部分，是展示科研人员科研能力的一项重要指标。

图 23 学者的专利与基金项目展示

人才发展预测也是当前学术搜索产品的特色功能，如图 24 所示。通过分析科研人员的现有成就，并与同领域人才进行对比分析，预测其未来的发展路径，便于对人才进行有针对性地培养和管理。

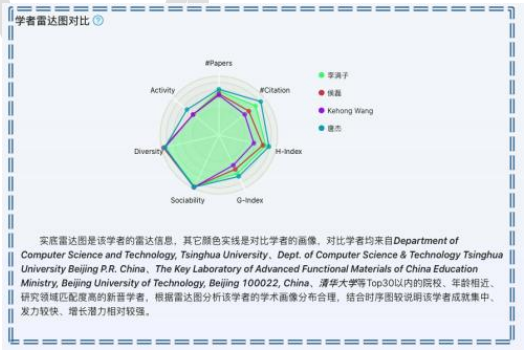


图 24 学者未来发展成就预测

4.7.6 知识图谱

知识图谱是通过将应用数学、图形学、信息可视化技术、信息科学等学科的理论与方法与计量学引文分析、共现分析等方法结合，并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构，把复杂的知识领域通过数据挖掘、信息处理、知识计量和图形绘制而显示出来，揭示知识领域的动态发展规律。

目前，一些学术搜索产品的知识图谱技术应用还较简单。例如，微软学术/必应学术通过**主题标签**展示各学科领域主题页面、子领域主题的详细页面。中国知网支持**智能输入提示、平面式分类导航、个性资源分类导航。**Web of Science

提供**增补关键词**功能，除了原文作者提供的关键词，而 ISI 在处理时加上了增补关键词字段 Keyword Plus。增补的关键词从原文的参考文献的篇名中选择有检索意义的词和短语，与作者自己的关键词对照后，将缺少的关键词列入。

基于知识图谱技术，百度学术产品支持**多维度论文推荐**，系统从经典论文、领域最新发表、综述论文、学位论文 4 个维度进行论文推荐帮助用户更详细的了解研究进展。此外，该产品还支持**关键词推荐，**根据输入的研究方向的关键词，找到有交叉研究的其他研究关键词进行推荐。

AMiner 提供了**技术分析与预测**功能。该产品以十余年积累的学者及机构命名排歧、主题聚类、隐含关联关系挖掘、知识图谱等关键技术为基础，提供系统全面的技术发展趋势分析、前沿技术预见、技术成熟度分析等科技情报分析挖掘工具及共性关键技术支撑。

此外，基于知识图谱，AMiner 产品还支持**论文溯源分析**，即通过检索、阅读、构图、推理等，构建论文演变图。

- ✓ 检索：从 AMiner 平台数据源中检索论文的直接引用与间接引用；
- ✓ 阅读：阅读论文及其引用，抽取特征并聚焦于最重要的一部分资料；
- ✓ 构图：分析引用之间的关系并将其作图于一棵树中；
- ✓ 推理：尝试从引用关系中学习它们相互作用的方式。

4.7.7 可视化分析

可视化分析主要应用于海量数据关联分析，可辅助人工操作将数据进行关联分析，并做出完整的分析图表。目前，学术搜索产品都非常重视并积极推出可视化分析功能。

Science.gov **可视化展示**了文献主题类别分布。Scopus 将搜索结果按年份、来源、作者、国家、文献类型、主题等进行**可视化分析**展示。

百度学术将可视化分析应用于**研究点分析**和**开题分析。**在研究点分析中，系统抽取出与所检索文献最相关的多个研究点进行深度分析，用户点击后即可查看可视化分析。在开题分析中，将输入的功能点从研究走势、关联研究、学科渗透、

相关学者、相关机构 5 个方面进行可视化分析，帮助用户全方位了解该研究方向的历史研究进展、交叉学科及重点研究学者&机构等。

AMiner 将**学者信息**进行**可视化展现**，具体包含了学者研究兴趣变化、合作者网络、相似学者、个人经历、以及发表的论文数量、引用量、发表期刊的影响力、学者的社交活跃度、研究兴趣跨度等做统计分析及可视化展示。

图 25 学者信息可视化展示

4.7.8 文献管理

文献管理是学术搜索产品的重要辅助功能，通常包括文献的引用、导出以及分享等。目前产品都有不同程度的文献管理功能。

Google Scholar 自动**分析和抽取引用**，系统从全文索引中构建一个**引文索引**，作为其服务的附加组件，并将其引用分析作为单独的结果显示出来。此附加可以对全文进行统计最佳匹配排名，还可以重新排列文档或对某些文档集进行分析和评估。自动参考文献的抽取和分析，也称为自动引文索引（ACI），在信息检索和传递中对用户特别有用。

微软学术/必应学术的**添加**功能，支持把论文添加到阅读列表、添加到引用、**分享论文**链接到 Facebook、twitter 等；Science.gov **管理功能**包括把文献加入收藏、打印、邮件发送、提醒等；BASE 系统支持**保存搜索结果**，并将文献输出（支持 RefWorks, EndNote,RIS, BibTex, MARC, RDF, RTF, JSON,YAML）；Scopus 可导入个人文献管理软件中；ScienceDirect 支持**导出** pdf 到目录管理软件（RefWorks, RIS, BibTex）。

中国知网支持**在线阅读、组合在线阅读、文献分析、多次查询结果一次性存盘导出、跨平台文献分享**等多项管理功能。Web of Science 支持将检索结果打印

输出、输出到电子邮件信箱、输出到目录管理软件，比如 Reference Manager，ProCite，EndNote 等。Semantic Scholar 导出引用、保存到个人图书馆、相似主题等。Semantic Scholar 文章可按 BitTex、EndNote、MLA、APA、Chicago 格式导出引用。

百度学术专门设立了**功能区**，每篇文献均在功能区提供引用、收藏功能。系统在文献功能区提供了单篇“引用”功能，用户可以根据所需格式（GB/T 7714、MLA、APA）进行选择；还将当前文献被引情况按年度进行**引用统计**，方便用户快速了解该文献在所属领域的影响力情况。同时，**支持批量引用**。支持 BibTex、EndNote、RefMan、Notefirst、NoteExpres 五种文献管理软件导入格式。在**收藏**功能中，系统支持将收藏论文按标签进行分类；支持批量导出；展示推荐文献；支持全文阅读&添加备注等。

4.7.9 学术资讯推送

学术资讯推送可以方便用户及时了解学术科研动态，并非产品必备功能。一般情况下，用户需要先行订阅才能收到系统推送。

BASE、中国知网和百度学术等产品都提供**订阅推送**功能。BASE 支持发送电子邮件订阅。百度学术目前支持对关键词进行订阅，当相关且符合订阅设置的新研究成果出现时，会自动推送到用户，推送频率为每周 2~3 次，推送包括系统消息推送、邮箱推送、微信推送三种推送方式。

AMiner 系统支持**精准推送**。基于高质量语义分析和内容生成技术，系统结合专家画像，在 AMiner 平台覆盖的全球 1.36 亿位学者中帮助期刊在交叉领域和新兴热点领域寻找国内外合适的约稿对象和审稿人，从而进行精准推送。

4.7.10 用户个人档案

学术搜索产品通常为已注册或付费的用户提供个人档案模块。用户可以在其中进行保存搜索记录、管理个人账号资料等个性化操作。例如，微软/必应学术的**用户个人档案**。以个人档案为基础，微软/必应学术提供了可视化服务，个性化搜索等功能。用户可以查看学术地图、作者关系图、引文关系图等。可编辑个

人的基本信息，如姓名、所属机构、头像及个人主页；论文的基本信息；上传论文；确认论文归属等。Web of Science 用户可自建**文献数据库，**并进行数据再处理。Semantic Scholar 支持用户将搜索结果文章保存到个人图书馆等。百度学术支持用户实名认证后自己管理个人主页，认证后可以进行编辑个人基本信息，增删或导出学术成果等**管理操作，**成果被引提醒等。

4.8 产品功能小结

综上所述，学术搜索产品的主要功能项可划分为以下几类：S-专家检索；N-网络关系分析；K-知识图谱；D-命名排歧；I-多源数据融合；V-可视化分析；P-学术资源推送；R-审稿人推荐；B-文献管理；E-学术指标评价。

学术搜索产品主要功能支持汇总如表 6 所示。

表 6 学术搜索产品主要功能一览

公司	产品名称	学科领域	主要功能									
			S	N	K	D	I	V	P	R	B	E
谷歌	**Google Scholar**	全学科	●	○	○	○	●	○	○	○	○	●
微软	**Academic Search**	全学科	●	●	○	●	●	●	○	○	○	●
宾夕法尼亚州立大学	**CiteSeerX**	计算机	●	○	○	●	●	○	●	○	●	●
ResearchGate	**ResearchGate**	全学科	●	●	○	○	●	○	●	○	●	●
爱思唯尔	**Scopus**	全学科	●	○	○	●	●	○	○	○	○	○
AI2	**Semantic Scholar**	全学科	●	○	○	○	●	●	●	○	●	●
特里尔大学	**DBLP**	计算机	○	○	○	○	○	○	○	○	●	○
Nature	**Digital Science**	全学科	●	○	●	○	●	○	○	●	●	●
美国能源部	**Science.gov**	全学科	○	○	○	○	●	○	●	○	●	○
比勒费尔德大学	**BASE**	全学科	○	○	○	○	●	○	○	○	●	○
英 Open 大学	**CORE**	Open	○	○	○	○	●	○	○	○	●	○
汤森路透	**Web of Science**	全学科	●	○	○	○	●	●	●	○	●	●
爱思唯尔	**ScienceDirect**	全学科	●	○	○	○	●	○	●	●	●	○

中国知网	**CNKI**	全学科	●	○	●	○	●	●	○	○	●	●
百度	**百度学术**	全学科	●	○	○	○	●	○	●	○	●	○
清华大学	**AMiner**	全学科	●	●	●	●	●	●	●	●	●	●

5

趨勢篇

学术搜索领域的技术发展趋势与人工智能技术的突破性研究息息相关。总体而言，在学术搜索领域，人工智能中的突破性研究与学术搜索引擎的结合将是未来学术搜索技术发展的主要趋势。其中，使用自然语言处理技术来优化学术搜索引擎的使用感受，基于深度学习而实现系统对论文内容的理解，消除作者或语义歧义，用计算机模拟人的视觉来描述图像的内容，并根据内容描述的特征从海量的图像中找出感兴趣的目标图像等都将实现得更加完美。

未来，基于人工智能技术的学术搜索不仅可以分析研究论文，辅助学术科研工作，而且将承担挖掘科技竞合情报，助推企业、行业乃至产业的创新变革的重要角色。

5.1 AI 学术搜索的技术发展趋势

领域技术分析可以为学者未来确定科研方向及科技计划提供参考。基于 AMiner 领域技术分析系统（http://trend.aminer.cn），我们对 AI 学术搜索领域国内外顶级期刊会议中发表的大量论文和学者信息进行深入挖掘，探索分析了 AI 学术搜索技术发展趋势、各个国家研究热度趋势以及科研跨国合作情况。

技术发展趋势分析描述了技术的出现、变迁过程，可以帮助研究人员理解技术领域的研究历史和现状，快速识别研究的前沿热点问题。AI 学术搜索技术发展趋势分析如图 26 所示。图中每条色带表示一个话题，其宽度表示该话题在当年的热度，与当年该话题的论文数量呈正相关，每一年份中按照其热度由高到低进行排序。通过技术发展趋势分析可以发现当前 10 大热点研究话题是：Information Retrieval（信息检索）、Search Algorithm（搜索算法）、Image Retrieval（图像检索）、Data Mining（数据挖掘）、Machine Learning（机器学习）、Search Space（搜索空间）、Web Search Engine（网页搜寻引擎）、Feature Extraction（特征抽取）、Indexation（索引编制）、Natural Language Processing（自然语言处理）。

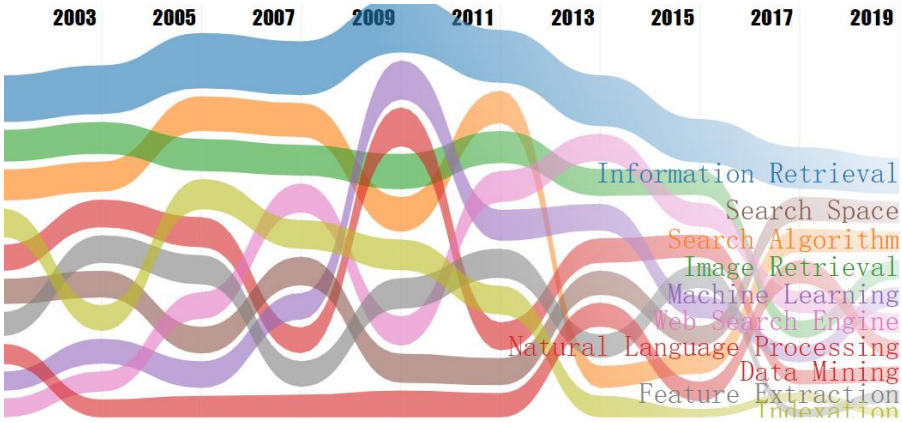


图 26 AI 学术搜索技术发展趋势图

通过抽取学者近 10 年发表论文数据的关键词，进行聚类分析以及词频统计等算法，绘制出学者们的研究兴趣词云图，如图 27 AI 学术搜索技术研究热点词云图所示。从中可以看出，领域学者的研究兴趣主要集中在 Information Retrieval（信息检索）、Search Algorithm（搜索算法）、Image Retrieval（图像检索）、Data Mining（数据挖掘）和 Machine Learning（机器学习）等方向。

图 27 AI 学术搜索技术研究热点词云图

5.2 关于 AI 学术搜索产品性能升级的建议

目前，学术搜索产品正在不断地优化技术和产品性能。通过我们的调研发现，用户对 AI 学术搜索存在很多期待。据此，我们对 AI 学术搜索产品的未来升级提出以下建议，供产品开发者和服务者参考。

一是**扩展现有产品的学术资源覆盖**。由于知识产权或合作权限等限制，目前产品无法做到完全覆盖所有学术资源。而用户则希望搜索出来的结果全面准确。因此，建议产品方需要开拓更多合作渠道获取学术资源为用户服务。其中，会议论文资源建议更多增加。

二是**升级现有产品的文献检索功能**。在现有搜索功能基础上，**建议增加更多搜索域**，例如，按学校/机构名称查询、按学术主题/大类/研究分支进行论文检索。在搜索结果中，建议展示出**研究领域代表性论文、重要文献**。关于论文引用，建议**增加论文第一作者/通讯作者的引用量**，并且能够区分出是否为自引。

三是**将结果更多地进行可视化展现。**用户希望将搜索结果进行统计，以数据图表等可视化方式将结果展示出来。

四是**增加领域作者和技术的更多内容**。关于作者，建议提供学者详细信息描述、研究方向、学者动态以及相似学者关联等；关于技术研究，建议提供技术研究动态和研究热点、学术前沿技术、最新成果、趋势预测等。

五是**优化辅助功能**。现有推荐功能受到用户喜欢。建议今后推荐相关论文、最新论文或作者时，能够提高论文和学者推荐的**准确性**；根据用户偏好或以往搜索历史等进行**更多的个性化推荐**。此外，建议增加**自动翻译、分类收藏**论文等辅助功能。

六是**逐渐实现更多的 AI 突破**，减轻用户的科研负担。我们发现，用户对学术搜索产品的“智能”特色期待较高。用户不仅希望通过学术搜索产品能够自动识别关键论文、进行关键词联想、自动生成 bibliography 文件，而且希望产品能够帮助快速梳理某一领域的研究脉络、文献思想传承，自动生成文献综述，进而自动对论文进行归纳总结，并且展示出论文用途，甚至能回答自己的任何学术问题，成为自己科研工作的“阿法狗”。

5.3 AI 学术搜索的前沿技术热点

根据 AI 学术搜索领域近十年的相关论文，利用大数据分析、机器学习、人工智能等技术手段，建立算法模型及研发 demo 系统，分析挖掘出该领域的技术发展热点，绘制出技术预见图（如图 28 所示）。

技术预见图中点的大小表示该技术的热点（主要由相关论文数量的多少决定，相关论文越多，热度越高、点越大），各技术之间的连线表示该技术关键词与其他热词在该领域论文中共同出现且频率较高，选取与该词共现频率最高的前 3 个词进行连线后形成。

根据 AI 学术搜索领域技术预见图，可以发现领域相关度最高的技术主要为：data mining（数据挖掘）、support vector（支持向量）、social network（社交网络）、machine learning（机器学习）、feature extraction（特征抽取）、computer mediated communication（计算机中介传播）、collaborative filtering（协同过滤）以及 search engine（搜索引擎）等。

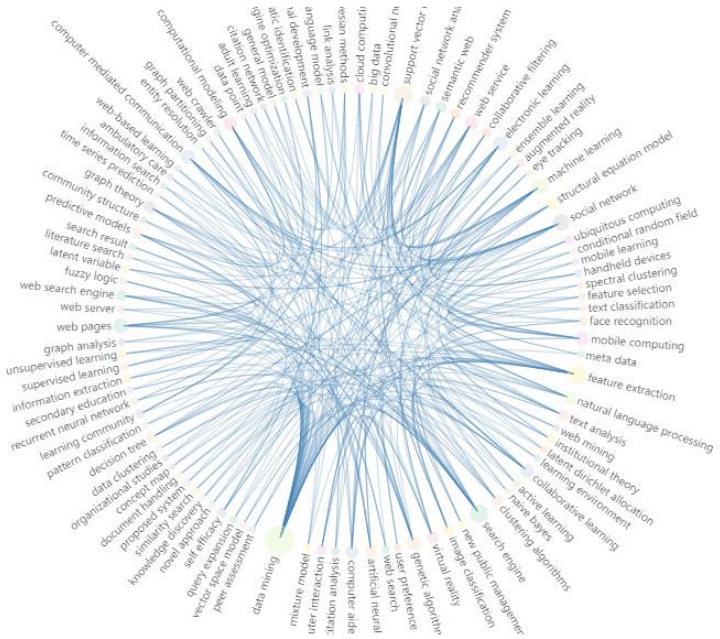


图 28 AI 学术搜索技术预见图

技术前沿度分析可以列出相关的主要技术关键词，以及该技术历年的变化趋势（论文发表数量变化趋势）。从技术前沿度分析图（如图 29 所示）可以看出，未来该领域的 12 个技术热点分别为：cloud computing（云计算）、big data（大数据）、support vector machine（支持向量机）、social network analysis（社交网络分析）、semantic web（语义网络）、web service（网络服务）、convolutional neutral network（卷积神经网络）、recommender system（推荐系统）、electronic learning（电子学习）、collaborative filtering（协同过滤）、ubiquitous computing（普适计算）、social network（社交网络）等。

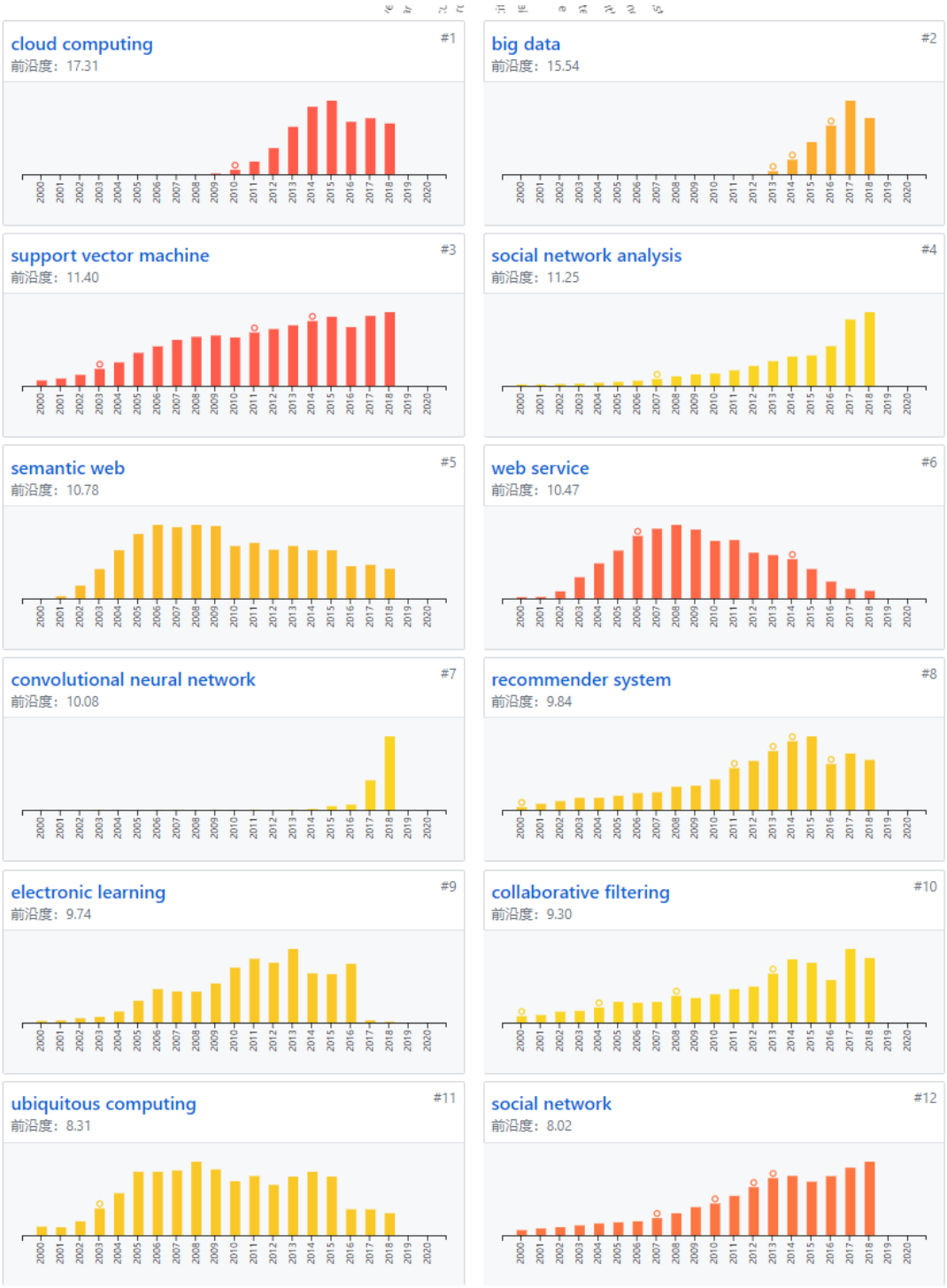


图 29 AI 学术搜索技术前沿度

5.4 AI 学术搜索的未来

随着 AI 和 IT 技术在学术领域的广泛和深入应用，学术搜索的未来发展将实现具有推理、具有可解释和具有认知的智能系统（如图 30），具有知识和算法双

引擎驱动的认知计算能力，呈现出**多源异构知识融合化、知识加工设施化和自动化、智能挖掘算法高度精准化、科技应用智能化和创新化**四大趋势。

1.多源异构知识融合化

未来，来自开放网络的数字学术资源将在来源、规模和种类等方面不断增多。这些学术资源涵盖的不仅是各个学科领域的期刊、论文、专利、会议、专家、机构等实体，还将覆盖到图形、表格、标准、年鉴、工具书、科学知识、报纸、新闻资讯、项目、评价甚至其他学术搜索引擎等更加多元化的学术资源。学术搜索平台需要对这些多源异构的资源进行整合分析、消除矛盾和歧义、融合并且获得新知识，从而为用户提供更多有价值的学术内容。

2.知识加工设施化和自动化

人工智能赋能的学术搜索系统涉及到许多知识的加工问题。通过科技知识表示、推理和计算，学术搜索引擎对存在于互联网上的各类学术资源进行二次提炼整合加工，在保持知识的原汁原味的基础上进行知识编撰和整理，再以图形的方式向用户返回经过加工和推理的知识，帮助用户在整个学术领域中确定相关性最强的科研信息。整个知识加工过程庞大繁杂，将越来越多地借助于 AI 系统设施和知识引擎技术来实现，以确保知识库的质量和效率。

3.智能挖掘算法高度精准化

通过先进的智能挖掘算法，AI 学术搜索引擎对知识有一定的理解和处理能力，可以智能解析和推理用户的搜索请求、搜索意图。或者，系统将允许用户使用自然语言通过交互式提问进行信息检索，并且通过用户搜索行为，识别用户的信息需求偏好；根据用户对搜索结果的反馈，调整搜索策略，提高搜索引擎的查准率。

支撑智能语义搜索、深度问答系统这些应用的核心技术正是知识图谱技术和搜索算法。这就需要更高精度、更高质量的搜索算法，以保证整个系统的搜索性能，方便用户获得领域的知识系统，全局化的知识结构。

4.科技应用智能化和创新化

随着学术搜索系统越来越智能化，学术搜索的应用范围将不仅局限于论文、专利、专家，以及期刊会议的搜索，未来将产生更多的创新式搜索应用和智能化科技服务领域，创造出更多的知识价值，实现学术上的创新和突破。

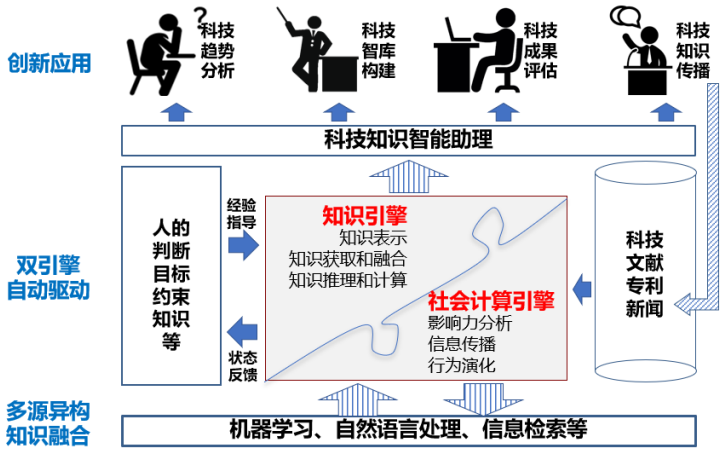


图 30 学术搜索的未来

总之，AI 赋能的学术搜索将不仅可以协助人们快速获取来自各个学科领域的学术资源，快速溯源某一领域发展脉络和探寻其研究新趋势新主题，快速发现领域高影响力的专家学者、参考文献、机构和国家，并且**将能够快速地将所有这些信息链接在一起，找到研究点之间的联系，形成一整套前沿研究的全貌，从而进行智能化的科技情报推荐和科技创新预测。**

在未来 10-20 年里，基于人工智能技术的学术搜索不仅可以分析研究论文，辅助学术科研工作，而且有能力**挖掘出聚集高知识性、高传递性和高效用的科技情报，在国家创新驱动发展战略实施中，将承担起挖掘科技竞合情报，助推企业、行业乃至产业的创新变革的重要角色。**通过深度参与情报感知、情报刻画和情报响应，人工智能将能够把科研之间的点点滴滴联系起来，在全面的时空视角下，进行推理假设，甚至进行实验验证给出答案，从而营造出一个健康、可持续的全球科技情报生态环境。

参考文献

[1] Arnab Sinha,,Zhihong Shen,,Yang Song,... & Kuansan Wang.(2015).An Overview of Microsoft Academic Service (MAS) and Applications..(eds.)World Wide Web(pp.)..

[2] G. W. Furnas,,S. Deerwester,,S. T. Dumais,... & K. E. Lochbaum.(1988).Information retrieval using a singular value decomposition model of latent semantic structure..(eds.)Research and development in information retrieval(pp.)..

[3] Academic search engines: a quantitative outlook. José Luis Ortega. Elsevier Science, Amsterdam, 2014, 198 págs. ISBN 978-1-84334-791-0 (print)

[4] Ballesteros, L., & Croft, W. B. (1998). Resolving ambiguity for cross-language retrieval. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 64–71).

[5] Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., & Li, H. (2007). Learning to rank: from pairwise approach to listwise approach. In Proceedings of the 24th international conference on Machine learning (pp. 129–136).

[6] Furnas G W, Deerwester S., Dumais S T, et al. Information retrieval using a singular value decomposition model of latent semantic structure[C]//Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1988: 465-480.

[7] Google Scholar: the ultimate guide， https://paperpile.com/g/google-scholar-guide/

[8] Haddaway, N. R., Collins, A. M., Coughlin, D., & Kirk, S. (2015). The role of Google Scholar in evidence reviews and its applicability to grey literature searching. PloS one, 10(9). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4574933/

[9] Jones K S. Index term weighting[J]. Information storage and retrieval, 1973, 9(11): 619-633.

[10] Li, J., Tang, J., Zhang, J. et al. Arnetminer: expertise oriented search using social networks. Front. Comput. Sci. China 2, 94–105 (2008).

[11] Manning C, Raghavan P, Schütze H. Introduction to information retrieval[J]. Natural Language Engineering, 2010, 16(1): 100-103.

[12] Nirkhi, Smita & Dharaskar, Rajiv & Thakare, V. M.. (2015). Authorship Identification using Generalized Features and Analysis of Computational Method. Transactions on Machine Learning and Artificial Intelligence. 10.14738/tmlai.32.1064.

[13] Ogawa Y, Morita T, Kobayashi K. A fuzzy document retrieval system using the keyword connection matrix and a learning method[J]. Fuzzy sets and systems, 1991, 39(2): 163-179.

[14] Robertson S E, Jones K S. Relevance weighting of search terms[J]. Journal of the American Society for Information science, 1976, 27(3): 129-146.

[15] Rosenfeld R. Two decades of statistical language modeling: Where do we go from here?[J]. Proceedings of the IEEE, 2000, 88(8): 1270-1278.

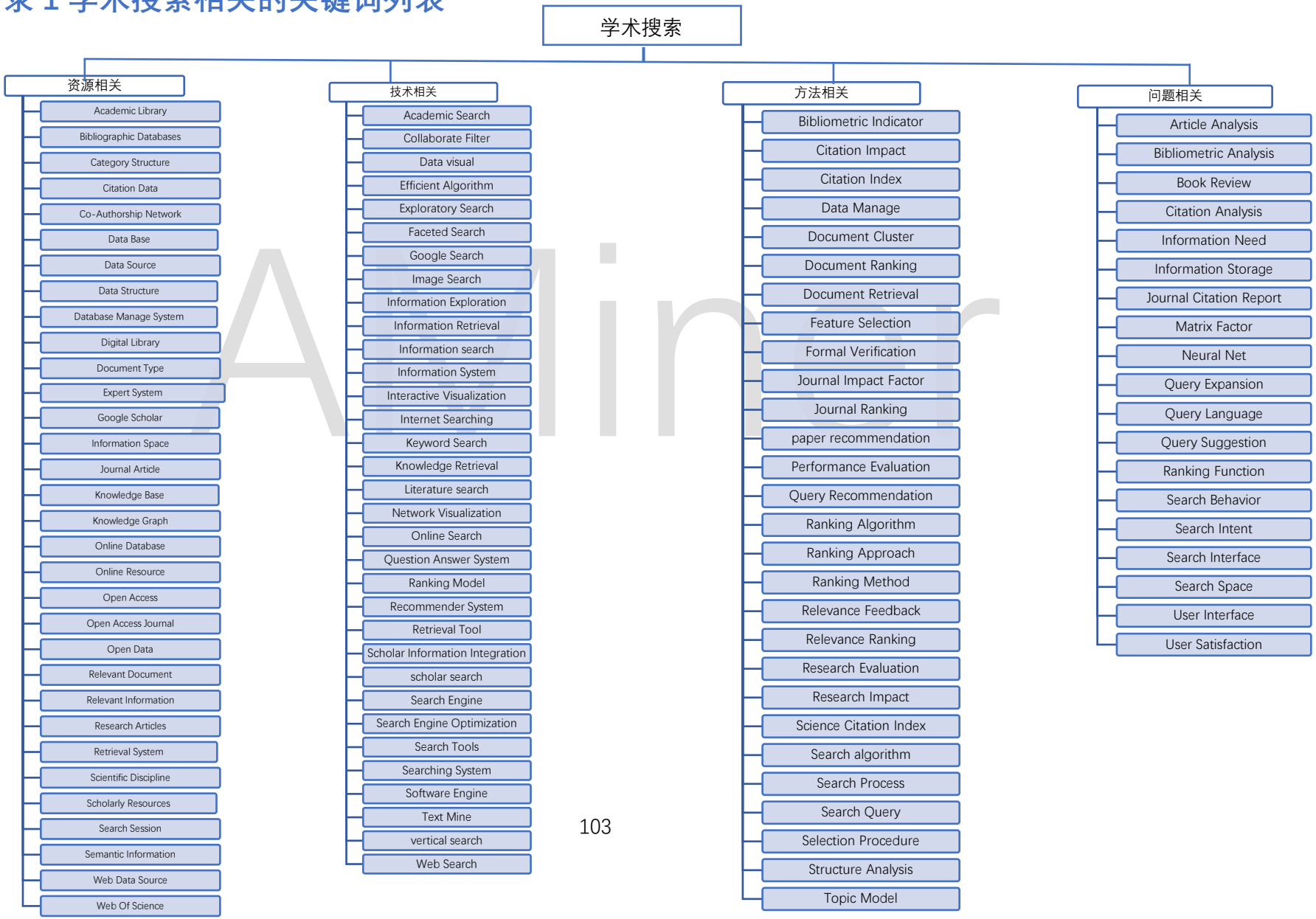
[16] Salton G, Fox E A, Wu H. Extended Boolean information retrieval[R]. Cornell University, 1982.

[17] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: extraction and mining of academic social networks. KDD.

[18] The top list of academic search engines， https://paperpile.com/g/academic-search-engines/

[19] Vine, R. (2006). Google scholar. Journal of the Medical Library Association, 94(1), 97. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1324783/
[20] Wang, K., Shen, Z., Huang, C., Wu, C. H., Dong, Y., & Kanakia, A. (2020). Microsoft Academic Graph: When experts are not enough. Quantitative Science Studies, 1(1), 396-413. https://www.mitpressjournals.org/doi/full/10.1162/qss_a_00021
[21] Wang, K., Shen, Z., Huang, C. Y., Wu, C. H., Eide, D., Dong, Y., ... & Rogahn, R. (2019). A Review of Microsoft Academic Services for Science of Science Studies. Frontiers in Big Data, 2, 45. https://www.frontiersin.org/articles/10.3389/fdata.2019.00045/full
[22] Zhang Y, Chen X. Explainable Recommendation: A Survey and New Perspectives[J]. 2018.
[23] Zhang, Y. , Zhang, F. , Yao, P. , & Tang, J. . (2018). Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop.. , 1002-1011.
[24] (2015).Academic Search Engines: A Quantitative Outlook. Online Information Review(3),.
[25] 《国家中长期科学和技术发展规划纲要（2006—2020 年）》中华人民共和国国务院，http://www.gov.cn/jrzg/2006-02/09/content_183787.htm
[26] 科技部发布国家“十二五”科学和技术发展规划，2011 年 07 月 13 日，http://www.gov.cn/gzdt/2011-07/13/content_1905915.htm
[27] 《2019 中国科技论文统计结果发布：从求数量到重质量 评价指标变化显著》，光明日报，http://www.gov.cn/shuju/2019-11/20/content_5453698.htm
[28] 朱雯,陈荣 & 孙济庆.(2019).多源数据的文献计量功能发展及其比较研究. 图书馆理论与实践(10),66-71.
[29] 李金忠,刘关俊,闫春钢 & 蒋昌俊.(2018).排序学习研究进展与展望. 自动化学报(08),1345-1369.
[30] 陈春玮.(2017).基于关联规则和神经网络分析的推荐系统的研究(硕士学位论文,杭州电子科技大学).https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201801&filename=1017133526.nh
[31] 孙建文.(2015).基于深度学习的中文文档检索的应用(硕士学位论文,吉林大学).https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201502&filename=1015591116.nh
[32] 孔维梁.(2013).协同过滤推荐系统关键问题研究(博士学位论文,华中师范大学).https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CDFD1214&filename=1014146932.nh
[33] 王静.(2012).基于关联规则的图书销售网站个性化推荐系统设计与实现(硕士学位论文,电子科技大学).https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201301&filename=1012472462.nh
[34] 李伟.(2007).基于关联规则 B2C 图书销售网站个性化推荐系统研究(硕士学位论文,对外经济贸易大学).https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD2007&filename=2007074858.nh

附录 1 学术搜索相关的关键词列表



附录 2 AI 学术搜索专家学者挖掘的来源期刊会议列表

刊物/会议全称	简称
AAAI Conference on Artificial Intelligence	AAAI
Annual Meeting of the Association for Computational Linguistics	ACL
ACM Conference on Computer and Communications Security	CCS
ACM CHI Conference on Human Factors in Computing Systems	CHI
ACM International Conference on Information and knowledge Management	CIKM
IEEE Communications Magazine	CM
International Conference on Computational Linguistics	COLING
ACM Conference on Computer-Supported Cooperative Work & Social Computing	CSCW
IEEE Conference on Computer Vision and Pattern Recognition	CVPR
Design Automation Conference	DAC
European Conference on Computer Vision	ECCV
Conference on Empirical Methods in Natural Language Processing	EMNLP
IEEE Annual Symposium on Foundations of Computer Science	FOCS
Symposium on Field Programmable Gate Arrays	FPGA
IEEE International Conference on Acoustics, Speech and Signal Processing	ICASSP
IEEE International Conference on Computer Vision	ICCV
International Conference on Learning Representations	ICLR
International Conference on Machine Learning	ICML
IEEE International Conference on Robotics and Automation	ICRA
IEEE Visualization Conference	IEEE VIS
International Joint Conference on Artificial Intelligence	IJCAI
IEEE Internet of Things Journal	IoT-J
IEEE/RSJ International Conference on Intelligent Robots and Systems	IROS
IEEE International Solid-State Circuits Conference	ISSCC
International Semantic Web Conference	ISWC
ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	KDD
International Conference on Principles of Knowledge Representation and Reasoning	KR
ACM International Conference on Multimedia	MM
ACM International Conference on Mobile Computing and Networking	MobiCom
North American Chapter of the Association for Computational Linguistics	NAACL
Annual Conference on Neural Information Processing Systems	NeurIPS
USENIX Symposium on Operating Systems Design and Implementation	OSDI
ACM Recommender Systems	RecSys
IEEE Symposium on Security and Privacy	S&P

ACM SIGCOMM Conference	SIGCOMM
ACM SIGGRAPH Conference	SIGGRAPH
International ACM SIGIR Conference on Research and Development in Information Retrieval	SIGIR
ACM SIGMOD International Conference on Management of Data	SIGMOD
ACM Symposium on Operating Systems Principles	SOSP
ACM Symposium on Theory of Computing	STOC
IEEE Transactions on Audio, Speech, and Language Processing	TASLP
IEEE Transactions on Information Forensics and Security	TIFS
IEEE Transactions on Information Theory	TIT
IEEE Transactions on Knowledge and Data Engineering	TKDE
IEEE Transactions on Multimedia	TMM
ACM Transactions on Graphics	TOG
ACM Transactions on Information Systems	TOIS
IEEE Transactions on Robotics	TR
IEEE Transactions on Visualization & Computer Graphics	TVCG
IEEE Transactions on Very Large Scale Integration (VLSI) Systems	TVLSI
IEEE Transactions on Wireless Communications	TWC
ACM International Conference on Ubiquitous Computing	UbiComp
USENIX Security Symposium	USS
International Conference on Very Large Data Bases	VLDB
Journal of Web Semantics	WS
ACM International Conference on Web Search and Data Mining	WSDM
International World Wide Web Conference	WWW
Neural Networks	
Information Fusion	
IEEE Transactions on Fuzzy Systems	
IEEE Transactions on Image Processing	
IEEE Computational Intelligence Magazine	
International Journal of Computer Vision	
IEEE Transactions on Evolutionary Computation	
IEEE Transactions on Neural Networks and Learning Systems	
IEEE Transactions on Pattern Analysis and Machine Intelligence	
IEEE Internet of Things Journal	
IEEE Transactions on Cloud Computing	
IEEE Communications Surveys and Tutorials	

附录 3 学术搜索领域国内外重要奖项

国内重要奖项	国外重要奖项
国家科学技术进步奖 国家教委科技进步奖 国家自然科学奖 高校自然科学奖 国家技术发明奖 中国科学院自然科学奖	Eugene Garfield Information Sciences Pioneer Award Citation Laureates Highly Cited Researchers Publons Peer Review Awards

版权声明

AMiner 研究报告版权为 AMiner 团队独家所有，拥有唯一著作权。AMiner 咨询产品是 AMiner 团队的研究与统计成果，其性质是供用户内部参考的资料。

AMiner 研究报告提供给订阅用户使用，仅限于用户内部使用。未获得 AMiner 团队授权，任何人和单位不得以任何方式在任何媒体上（包括互联网）公开发布、复制，且不得以任何方式将研究报告的内容提供给其他单位或个人使用。如引用、刊发，需注明出处为“AMiner.org”，且不得对本报告进行有悖原意的删节与修改。

AMiner 研究报告是基于 AMiner 团队及其研究员认可的研究资料，所有资料源自 AMiner 后台程序对大数据的自动分析得到，本研究报告仅作为参考，AMiner 团队不保证所分析得到的准确性和完整性，也不承担任何投资者因使用本产品与服务而产生的任何责任。

顾 问：李涓子 唐 杰

主 编：刘 佳

编 写：张 淼 高 洁

数 据：赵慧军

封面设计：张 淼

更多 AI TR 系列报告，敬请查看官方网址 http://reports.aminer.cn/

如果您对科技写作有兴趣和经验，或有商务合作需求，
欢迎联系我们 reports@aminer.cn

关注“学术头条”并回复
“学术搜索”下载报告。

扫描下方二维码
下载报告全文