pyBibX - 一款借助人工智能工具进行文献计量学与科学计量分析的 Python 库

Valdecy
P
ereiravaldecy.pereira
*@gmail.com*弗卢米嫩塞联邦大学
Marcio Pereira
B
asilio*marciopbasilio*
*@gmail.com*弗卢米嫩塞联邦大学
Carlos Henrique Tarjano
Santos*tesserato@hotmail.com*弗
卢米嫩塞联邦大学

摘要

背景：文献计量分析和科学计量分析为理解复杂的研究领域，以及不同学术学科间的协作动态，提供了极具价值的视角。目的：本文介绍 pyBibX，这是一款 Python 库，旨在利用 Scopus、Web of Science 和 PubMed 的原始数据文件开展全面的文献计量分析和科学计量分析，并将最先进的人工智能（AI）功能无缝集成到核心功能中。方法：该库可执行全面的探索性数据分析（EDA），并通过美观的图形展示结果。它还集成了网络分析功能，包括引文分析、协作分析和相似性分析。此外，该库还具备人工智能功能，包括嵌入向量、主题建模、文本摘要及其他通用自然语言处理任务，并使用 Sentence-BERT、BERTopic、BERT、ChatGPT 和 PEGASUS 等模型。研究发现：作为示例，我们分析了 184 篇与“多准则决策分析”相关的文档，这些文档发表于 1984年至2023年。探索性数据分析凸显了人们对决策和模糊逻辑方法日益浓厚的兴趣。随后，网络分析进一步强调了核心作者和洲内协作的重要性，并发现加拿大和中国是重要的协作枢纽。最后，人工智能分析识别出两个主要主题，并突显了 ChatGPT 在文本摘要方面的优势’。事实证明，它还是一种不可或缺的结果解读工具，因为我们的库支持研究人员就文献计量分析结果向 ChatGPT 提问。尽管如此，由于数据库之间存在不一致，数据同质性仍是一项严峻挑战。意义：pyBibX 是首个集成前沿人工智能功能、用于分析科学出版物的应用程序，使研究人员能够更有效地考察和解读分析结果。pyBibX 可从以下网址免费获取：https://bit.ly/442wD5z

关键词：文献计量学、科学计量学、网络分析、人工智能、ChatGPT

1. 引言

文献计量分析和科学计量分析可用于分析和量化特定文献集合的特征。这两种方法都采用数学与统计技术；然而，文献计量分析侧重于书籍和书面交流形式（OTLET，1934年；PRITCHARD，1969年；FAIRTHORNE，1969年），而另一种方法则专门研究科学研究集合（SENGUPTA，1992）。总体而言，两种方法在研究对象和分析技术方面可能有所重叠；因此，我们可以认为，根据研究情境，这两个术语可以互换使用。

针对特定文集，可以采用一些定量指标进行评估，例如文档总数、平均被引次数、平均期刊影响因子、平均 h指数、每位作者的平均出版物数量、科研机构总数等。此外，还可以通过引文网络、共被引网络和人工智能分析识别科学文献中的模式与趋势（SRISAWAD & LERTSITTIPHAN，2021年；BIRCAN & SALAH，2022年）。

分析结果可用于评估研究人员、期刊和科研机构的生产力与影响力，并识别研究领域的优势与不足。举例来说，这些结果可以帮助决策者确定特定研究领域中最具影响力的研究人员或机构、趋势或新兴趋势，以及研究人员与潜在合作者之间的合作模式。借助这些信息，决策者可以作出资助决策，并制定科研政策与战略。

不可否认，与任何其他技术一样，尽管这些方法的应用乍看之下颇具前景，但它们也存在局限性和潜在问题。主要问题包括：a) 误用——定量指标可能被过度解读，从而导致评估不准确。b) 引文数据——分析通常依赖引文数据，而这些数据可能存在偏差且不完整。c) 选择偏差——所分析的文献集合可能无法代表某一领域的全部研究成果。d) 引文偏差——一些研究人员可能通过自引其作品或参与引文联盟来操纵定量指标。这种做法可能会歪曲研究质量或重要性的真实信息（HUG，2003；AHMED 等，2012年；MIRIAN 等，2021年）。

上述所有问题都很严重，但这种类型的分析所带来的益处可能超过这些问题造成的不利影响。鉴于此，经过多年的发展，许多研究前沿应运而生，旨在推广并促进对文献计量学的审慎使用。这些研究前沿促成了两个相继发生的事件，为我们如今的文献研究格局奠定了基础：科学数据库的创建，以及能够处理这些详尽科学文献数据的软件的开发（BELLIS，2009年；BOURNE 等，2003）。

各种数据库不断涌现；然而，出于复杂的原因，一些数据库脱颖而出，例如 Web of Science（Web of Science）、Scopus 和 PubMed。Web of Science 是由 Clarivirate 运营的付费访问平台，提供丰富的科学文献资源，涵盖科学、社会科学和人文科学等广泛学科。Scopus 是由 Elsevier 运营的付费访问平台，也涵盖科学、技术、医学、社会科学和人文科学等广泛学科。最后，PubMed 是由美国国立卫生研究院（NIH）运营的免费数据库，收录全球生物医学文献（FALAGAS 等，2008年）。

上述数据库支持通过查询或搜索字符串来检索和发现相关的文章、书籍、会议论文集及其他资料。搜索结果汇编在一个文件中，其中包含每篇文档的相关信息（元数据），例如作者姓名、机构、摘要等。当然，同一篇文档在不同数据库中的元数据可能会相互冲突。此外，值得注意的是，各数据库可能会提供独有类型的元数据，例如文档的接收日期和时间（KOKOL & VOŠNER，2018年）。

许多文献计量软件都能处理数据库生成的文件，这些文件可能包含大量文档，具体数量取决于检索式。文献计量软件的主要目标是提供准确、可靠且透明的指标，用于评估研究影响力和生产力，因此必须仔细关注数据质量和方法论。

值得注意的是，近年来人工智能极大地推动了众多知识领域的发展，包括自然语言处理（NLP）。这些进展得益于深度学习模型，换言之，得益于能够从大量数据中学习的神经网络模型。例如，人工智能和文献计量学可以结合使用，以识别研究主题聚类（主题建模）、总结相关概念及其关系（文本摘要），并分析大量文献计量结果（使用 ChatGPT 执行自然语言处理任务）。这些任务之所以能够实现，是因为采用了嵌入技术，这种技术可将词语、短语或其他对象表示为高维向量空间中的向量。这些向量旨在捕捉其所表示对象的语义和上下文含义。将词语转换为向量，再将相同的向量用作深度学习模型的输入，就能开发出理解并生成类人语言的算法（DEVLIN 等，2019；ZHANG 等，2020；REIMERS 和 GUREVYCH，2019）。

因此，pyBibX 是一个由人工智能工具驱动的 Python 库，也是对文献计量分析的又一贡献。文献计量分析是科学研究中的重要工具，可帮助洞察研究趋势、识别重要研究人员，并评估科学出版物的重要影响。然而，传统的文献计量分析方法通常耗时较长，且需要大量人工工作。采用人工智能技术能够显著提升文献计量分析的准确性和效率。随着科学出版物呈指数级增长，我们需要更先进的文献计量分析技术，以应对所产生的海量数据。人工智能技术能够快速、准确且客观地处理大量数据，使研究人员得以获得原本难以获取的洞见。文献计量分析所获得的洞见可为科研资助决策提供参考，指导战略规划，并有助于制定促进科学研究的政策。因此，开发一款运用人工智能技术进行文献计量分析的工具，能够为科学界乃至整个社会带来显著益处。

本文分为五个部分，包括引言。第 2 节讨论 pyBibX 的主要目标，并将其功能与其他文献计量软件进行比较分析。第 3 节介绍本库使用的相关元数据，以及合并来自不同数据库的信息时面临的主要困难。第 4 节展示该库在探索性数据分析（EDA）、网络分析以及人工智能工具应用方面的能力，并讨论所获得的文献计量分析结果。最后，第 5 节给出总结性意见。

2. 相关工作

本节将对 pyBibX 的独特特性与其他知名文献计量工具进行比较评估。通过这种分析性对比，我们旨在确定 pyBibX 的相对效能与优势，从而凸显采用 pyBibX 的潜在益处；同时，我们也力求全面理解其在塑造文献计量分析未来格局方面所发挥的作用。

目前有许多文献计量软件可用于开展文献计量分析和可视化。表01列出了一些优秀的示例。

表01 – 文献计量工具

Name	数据源	环境	版本
pyBibX	PubMed、Scopus、Web of Science	Python 库	2.9.3（2023年4月16日）
Bibexcel（Persson 等人，2009年）	WoS（需要数据适配）	Windows（Excel）	1.6.4（2017年）
Bibliometrix（Aria 和 Cuccurullo，2017年）	Cochrane、Dimensions, PubMed、Scopus、Web of Science	R 软件包	4.1.2（2023年3月7日）
BiblioTools（Grauwin 和 Jensen，2011年）	Scopus、Web of Science	Python 库	3.2（2018年3月1日）
CiteSpace（CHEN，2006年）	Wos	基于Java	6.2.R2（2023年3月26日）
CitNetExplorer（Van Eck 和 Waltman，2014年）	Scopus、Web of Science	基于Java	1.0.0（2014年3月10日）
CRExplorer（Thor 等，2016年）	Scopus、Web of Science	基于Java	1.9（2018年7月16日）
Litstudy（Heldens等，2022年）	Scopus及其他数据库	Python 库	1.0.5（2023年3月28日）
Metaknowledge（McLevey，2017年）	PubMed、Scopus、Web of Scien及其他	Python 库	3.4.1 (2021年11月3日)
Publish or Perish（Harzing，2010）	PubMed、Scopus、Web of Scien及其他人	Windows	8.2.3944（2022年5月12日）
Sci2 Tool（Börner 等，2003）	Scopus、Web of Science 及其他	Windows、macOS、Linux	1.3.0（2020年6月24日）
Scientopy（Ruiz-Rosero 等，2019年）	Scopus、Web of Science	Python 库	2.1.2（2023年3月24日）
SciMat（Cobo 等人，2012年）	Scopus、Web of Science	Windows、macOS、Linux	1.1.04（2016年7月12日）
Tethne（Peirson 等人，2016年）	Scopus、Web of Science	Python 库	0.8（2016年6月21日）
VOSviewer（Van Eck 和 Waltman，2010）	Dimensions、PubMed, Scopus、Web of Science	Windows、macOS、Linux	1.6.19（2023年1月23日）

表01比较了15种文献计量分析工具，展示了它们各自独有及共有的功能。除名称列外，表中还包含以下列：数据源、环境和版本。

a) 数据源：每种工具或库都使用特定的数据来源，其中最常见的是 PubMed、Scopus 和 Web of Science。Bibliometrix、LitStudy、Metaknowledge、Publish or Perish 和 VOSviewer 将兼容范围扩展至 PubMed、Dimensions、Cochrane 等其他数据库。值得注意的是，Bibexcel 需要进行调整才能与 Web of Science 数据库配合使用。

b) 环境：此列重点介绍每种工具所运行的环境或平台。许多工具（如 pyBibX 和 BiblioTools）以 Python 库的形式运行，而其他工具（如 CiteSpace 和 CitNetExplorer）则是基于 Java 的应用程序。Bibexcel 和 Publish or Perish 仅适用于 Windows，Bibliometrix 则以 R 软件包的形式提供。同时，Sci2 Tool、SciMat 和 VOSviewer 支持 Windows、macOS 和 Linux 多个平台。最终，环境的选择取决于用户偏好、可访问性以及具体的项目需求。

c) 版本：此列列出了各工具的版本号和发行年份。近期更新的工具，如 pyBibX（2.9.3，2023年4月16日）、Bibliometrix（4.1.2，2023年3月7日）和 CiteSpace（6.2.R2，2023年3月26日），更有可能包含最新功能和改进。另一方面，较旧的工具，如 Bibexcel（1.6.4，2017年）和 Tethne（0.8，2016年6月21日），可能没有提供最新的改进。版本号和更新至关重要，因为它们表明开发者正积极发布更新，以优化功能、修复错误或引入新功能，确保工具始终兼容现代文献计量分析技术。

表01所列的文献计量工具在数据源、环境和版本列中呈现出一系列相似之处与差异。尽管存在差异，这些文献计量工具能够满足

多样化的研究需求与偏好，为研究人员提供广泛的文献计量分析和可视化选项，以满足其需要。

表02对表01中介绍的文献计量工具进行了深入比较，重点考察其在探索性数据分析、网络分析和人工智能方面的能力。探索性数据分析能力涵盖工具进行初步数据分析、识别模式与趋势以及实现数据可视化的能力，帮助研究人员发现洞见并提出假设。网络分析能力体现了工具考察并可视化作者、引文和关键词等实体之间复杂关系的水平，从而帮助研究人员解读数据结构及其动态变化。人工智能能力则运用深度学习技术，如主题建模、嵌入向量、文本摘要和通用自然语言处理任务，提升工具的效能。

表02 – 主要特性

Name																	
pyBibX	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Bibexcel		X								X		X					
Bibliometrix	X	X	X		X	X	X		X	X	X	X	X				
BiblioTools		X								X		X					
CiteSpace										X	X	X	X				
CitNetExplorer										X	X	X	X				
CRExplorer										X	X	X					
Litstudy		X							X	X	X	X	X	X			
Metaknowledge										X	X	X	X				
Publish or Perish		X															
Sci2 Tool		X								X	X	X	X				
Scientopy	X	X	X			X			X	X	X	X	X	X			
SciMat		X							X	X	X	X	X				
Tethne										X	X	X	X	X			
VOSviewer										X	X	X	X				

在表02中，这些工具最常见的共同功能包括引文分析、合作分析和世界合作分析，许多工具还支持相似性分析和主题建模。pyBibX 和 Scientopy 都提供了广泛的功能，涵盖表02中的大多数功能。CiteSpace、CitNetExplorer、CRExplorer、Metaknowledge、Tethne 和 VOSviewer 主要侧重于引文相关分析，其他功能较少。Bibexcel、BiblioTools、Litstudy、Sci2 Tool 和 SciMat 专注于数据处理和/或探索性数据分析，其中一些还提供引文、协作和主题分析。R包工具：Bibliometrix 是列表中唯一的 R 软件包，提供多种功能，包括数据处理、探索性数据分析和各种类型的分析。不过，pyBibX 凭借其丰富的功能脱颖而出，其中包括嵌入、文本摘要和 ChatGPT 等独特功能。

集成。另一方面，Publish or Perish 以极简设计著称，主要专注于探索性数据分析。

通过考察这两张表，我们可以发现一些见解和规律：pyBibX、BiblioTools、Litstudy、Metaknowledge、Scientopy 和 Tethne 都是功能各异的 Python 库。其中，pyBibX 和 Scientopy 的功能最为丰富，提供数据处理、探索性数据分析（EDA）以及多种可视化和分析技术，适合希望在单一工具中获得全面功能的研究人员。Litstudy 和 SciMat 也提供了相当丰富的功能，而 BiblioTools 和 Tethne 则更专注于引文和合作分析。CiteSpace、CitNetExplorer 和 CRExplorer 等基于 Java 的应用程序主要聚焦于引文、合作和相似性分析。它们的功能较为相似，但不如某些 Python 库全面。Bibexcel 和 Publish or Perish 是具有独特功能的基于 Windows 的工具。Bibexcel 要求对 WoS 纯文本进行数据适配，并依赖 Excel；而 Publish or Perish 支持更广泛的数据源。不过，与列表中的其他工具相比，这两款工具的功能都较为有限。Bibliometrix 是列表中唯一的 R 软件包。它提供了广泛的功能，包括数据处理、探索性数据分析（EDA）、可视化和多种类型的分析，是熟悉 R 的研究人员的灵活选择。此外，Bibliometrix、Litstudy、Metaknowledge 和 Publish or Perish 将支持扩展至多种数据源，因此对于使用不同数据库的研究人员而言更加灵活。Sci2 Tool、SciMat 和 VOSviewer 兼容 Windows、macOS 和 Linux 平台，并分别提供不同范围的功能。Sci2 Tool 和 SciMat 涵盖多种分析和可视化技术，而 VOSviewer 专注于可视化，不具备数据处理和探索性数据分析（EDA）功能。

将 pyBibX 与 ChatGPT 融入文献计量分析，为整个文献计量学领域带来了开创性的方法。ChatGPT 与 pyBibX 的结合简化了对复杂文献计量结果的解读与理解过程。ChatGPT 的对话功能让用户能够与分析软件顺畅互动，提出问题并获得相关答案。这种交流有助于厘清研究发现与模式，使研究人员无需费力地手动探索数据，便能轻松理解关键信息并识别趋势。

此外，将 ChatGPT 与 pyBibX 相结合，有助于通过弥合文献计量学与自然语言处理之间的差距来促进跨学科研究。这种协同作用推动开发新颖的方法与技术，进一步促进这两个领域的发展。因此，研究人员能够受益于更强大、更高效的文献计量工具，从而应对日益复杂的研究问题。这一强强联合有望重塑研究人员开展文献计量分析的方式，推动创新并发掘新的见解。

3. 科学数据库与元数据的利用

最常用的三种科学数据库是 Web of Science、Scopus 和 PubMed。这些数据库在科学文献覆盖范围方面虽有相似之处，但也各有独特之处。研究特定领域（如医学）的研究人员可能会发现 PubMed 最为相关，因为它对生物医学文献有深入的覆盖。另一方面，跨学科领域的研究人员可能会发现 Web of Science 或 Scopus 更有价值，因为它们覆盖了广泛的学科。

关于这些科学数据库提供的文档元数据，它们都是用于描述每份文档内容和上下文的一组结构化信息。不同数据库之间的结构化信息并不相同。这些差异可能源于多种因素，包括索引方法、错误或遗漏、不同的文档版本，以及数据库范围和侧重点。因此，被多个数据库收录的文档，其元数据可能存在差异。例如，总被引次数可能显示不同的总计数。表03展示了元数据项以及每项元数据所支持的技术。值得一提的是，除“参考文献”和“被引次数”外，所有元数据项都可以从这三个数据库中获取；PubMed不包含这两项元数据。

表03 – 数据库元数据

元数据	定义	探索性数据分析	网络分析	人工智能
摘要	.....的摘要 文档主要观点的总结以及 研究发现能让读者快速 文档的概述 目的和结论。	演化图，N- 克，投影， 词云		嵌入，文本 摘要、主题 建模
作者	贡献者们 到文档。	作者生产力 绘图、条形图、探索性数据分析 报告，桑基图 图表，	协作 分析	
作者 所属机构	组织的名称、地址及 组织所在国家 作者所属或任职的 机构。	探索性数据分析报告，柱状图 绘制，桑基图 图表	协作 分析，世界地图 协作 分析	
作者 关键词	关键词或短语列表 由作者选择的娱乐主管 总结主要主题或 文档的主题。	条形图，探索性数据分析 报告，演变 绘制 N 元语法， 投影，桑基图 图表，词云	邻接分析	嵌入
文档 标题	文档标题	演化图，N- 元组、投影， 词云		嵌入
文档 Type	文档类型（原创 研究论文、综述 文章等）。	探索性数据分析 报告		
期刊	文档发表于 的期刊 中。	条形图、演变 绘图，桑基图 图表		
关键词 Plus	附加关键词或 补充该内容的短语 作者选定的关键词和 提供更多背景信息以便了解 文档。	条形图、探索性数据分析 报告、演变 图、N元语法、 投影、桑基图 图表、词云	邻接分析	嵌入
语言	其中使用的语言 文章的撰写情况。	条形图，探索性数据分析 报告，桑基图 图表		
参考文献 （不在 PubMed）	文档中引用的参考文献列表 在文档中。	探索性数据分析 报告	相似性分析	
被引次数 （不在 PubMed)	其他人的次数 研究人员引用了 the 文档。	条形图，探索性数据分析 报告	引文分析， 相似性分析	

表03列出了我们的库所使用的元数据，包括其定义以及使用特定元数据的技术。这些技术分为三组：探索性数据分析、网络分析和人工智能。

表1中的探索性数据分析是指可用于分析文档元数据、以揭示模式、关系和趋势的各种技术与方法。表1列出的技术包括：

1) 作者生产力图：以图形方式呈现作者随时间变化的生产力，有助于追踪作者的产出并识别研究模式。2) 条形图：一种常见的图表类型，用于展示分类数据，尤其适合比较两个变量。我们的库提供多种条形图选项，将在第 3 节中讨论。从本质上说，每种选项都能以独特的视角呈现研究元数据，帮助人们深入了解研究领域的概况，例如最具影响力的作者、被引次数最多的文档，以及该领域最常用的关键词。通过在条形图中可视化这些数据点，研究人员能够更好地理解元数据中的趋势和模式，并据此指导研究工作。3) 探索性数据分析报告：汇总报告，提供元数据的描述性统计和可视化结果。它是数据分析中的关键步骤，有助于深入了解数据、识别模式与关系，并验证假设。报告会提供研究的时间跨度、国家和机构总数，以及来源和参考文献总数。报告还包括文档所用的语言、文档总数、文档类型，以及每位作者、每家机构、每个来源和每年的平均文档数。此外，报告还会提供作者人数、作者关键词数量和协作情况。最后，报告还会列出最高本地 h 指数和总被引次数，并给出每位作者、每家机构、每份文档和每个来源的平均引文数。4) 演化图：以图形方式呈现某个术语或主题的出现频率随时间的变化。这是可视化数据趋势和模式的实用工具，可用于识别行为或模式随时间发生的变化。5) N元语法：一种通过将文本划分为连续的 n 个词序列，从文本中提取词语模式的技术。6) 投影：将高维数据转换到低维空间，同时保留数据点之间结构和关系的方法。数据投影旨在降低高维数据的复杂性，并以更高效、更易理解的方式将其可视化。可使用两种算法之一：McInnes 等人（2018年）提出的统一流形近似与投影（UMAP），或截断奇异值分解（TSVD）。7) 桑基图：该图展示文档在七种元素（作者、国家、机构、来源、作者关键词、关键词增强词或语言）任意组合之间的流动。在桑基图中，线条的宽度代表研究成果的流动量，节点则代表不同的元素。8) 词云：以图形方式呈现文本中出现频率最高的词语，每个词语的字号反映其出现频率。词云有助于快速识别大型文本数据语料库中最常见的主题或话题。需要注意的是，词云存在局限性，因为它无法对文本数据的内容或语境进行详细分析。因此，为了更全面地理解数据，应将词云与其他文本分析技术结合使用。

表1中的“网络分析”列指用于分析元数据中不同要素之间关系和交互的技术，这些技术有助于理解所有研究的底层结构和动态变化。这些技术包括：

1）邻接分析：用于识别元数据中不同要素之间的邻近关系和连接，并将由此形成的网络结构可视化。2）合作分析：用于识别元数据中作者或机构的共现情况，并将它们之间的合作模式可视化。3）世界地图合作分析：用于标注元数据中作者或机构所在的位置，并在全球范围内将它们之间的合作模式可视化。

表1中的“人工智能”列指将深度学习算法应用于元数据以生成洞见并揭示模式的技术。这些技术包括：

1) 嵌入：此技术可以将词语或短语表示为高维空间中的向量，用于识别元数据中不同元素之间的相似性和关联。

2) 文本摘要：此技术用于生成元数据中文本数据的摘要，有助于快速理解文档的要点和主题。可以采用抽象式摘要和抽取式摘要。3) 主题建模：此技术用于识别元数据中潜在的主题或议题，有助于理解文档的结构和内容。4) 通用自然语言处理任务：此功能使用 ChatGPT 分析文献计量结果。可分析的结果包括探索性数据分析报告、词云、N元语法、演化图、桑基图、作者生产力图、条形图、引文分析、合作分析、相似性分析和世界地图合作分析。研究人员可以输入任何问题，以帮助理解文献计量结果。

如前所述，关于数据库中元数据的最后一个考量，是合并两个或更多数据库的信息。由于存在若干问题，合并来自 Scopus、Web of Science 和 PubMed 等不同来源的元数据可能颇具挑战。首先，这些数据库可能采用不同的元数据格式和标准，合并数据时因此可能产生不一致和错误。其次，这些数据库的覆盖范围和筛选标准可能不同，导致可供分析的数据有所差异。第三，这些数据库的准确性和完整性可能参差不齐，从而影响合并后数据的质量。最后，元数据中可能包含错误、重复项或缺失值，进一步增加了合并过程的复杂性。尽管我们的库允许手动编辑元数据元素，但这只能解决小规模问题，因为更正大量文档是一项难以想象的工作。

一种可行的解决方案是遵循特定顺序，并避免覆盖参考元数据。具体而言，先选择一个数据库（例如 PubMed）作为参考，用于合并其他数据库中的数据。随后将较新的元数据添加到参考数据中，但仅纳入参考数据中尚未包含的新信息。这样可以避免覆盖现有元数据，并确保合并数据集包含所有可用信息且没有重复项。

不过可以注意到，这一合并过程的顺序会保留先前的数据库，且合并顺序可能影响最终结果。例如，先合并 Web of Science 数据集再合并 Scopus 数据集，所得结果与先合并 Scopus 数据集再合并 Web of Science 数据集并不相同。在元数据合并过程中，选择从 Web of Science 或 Scopus 数据集开始，体现了研究人员对所选数据库的信心。随后添加另一个数据库的数据，可补充现有数据集。如果合并过程从 Web of Science 开始，Scopus 数据就会补充元数据，反之亦然。

尽管如此，采用这种方法仍然能够创建高质量且全面的数据集，充分利用多个数据库的优势，同时避免合并不一致且不完整的元数据所带来的挑战。

4. pyBibX – 文档集

为了展示我们的 Python 库 pyBibX 的强大功能，我们在演示中采用了一个范围较窄的搜索字符串。这样，读者便更容易理解我们着重介绍的概念。不过需要注意的是，我们的库能够处理复杂得多的搜索字符串，并为用户提供丰富的功能供其探索。我们鼓励用户访问 pyBibX 主页 https://pypi.org/project/pyBibX，进一步了解该库'的功能并探索其能力。

为了展示该库的功能，我们在 Scopus、Web of Science 和 PubMed 数据库中检索了标题包含以下查询词的文档 - "多准则决策分析" – 截至2023年1月1日已发表的文档。我们下载了各数据库可提供的所有文档信息。Scopus（mcda_scopus.bib）、Web of Science（mcda_wos.bib）和 PubMed（mcda_pubmed.txt）中获取的文件可通过以下链接访问：bit.ly/3XYK3eB。

搜索在 Scopus 中检索到 234 篇文档，在 Web of Science 中检索到 206 篇，在 PubMed 中检索到 10 篇。然而，Web of Science 中仅有 30 篇文档与 Scopus 中的不同；PubMed 中也仅有 1 篇文档与 Scopus 和 Web of Science 合并后的检索结果不同。按照 Scopus、Web of Science 和 PubMed 的顺序合并后，研究共包含 265 篇文档。

我们的库还可以根据各种标准筛选文档，包括文档类型、年份、来源、依据布拉德福定律（布拉德福，1934年）划分的来源、国家、语言以及是否包含摘要。布拉德福定律是书目数据中的一种统计模式，指出少数来源对特定研究领域的文献贡献显著。布拉德福定律通常以三区模型表示，其中核心来源（第一区）贡献的文档数量最多，其次是第二区中数量相对较少的文档，最后是第三区中的大量文档。

接下来，我们又增加了一步，对这 265 篇文档进行筛选，仅接受分类为“文章”“会议论文”“会议摘要”和“综述”的文档进行分析。我们没有对文档的年份、来源或所属国家作任何限制。不过，文档必须以英语撰写且包含摘要。经过筛选，最终数据集共包含 184 篇文档，将在以下各节中进行分析。

所有 184 篇文档都被赋予了从 0 到 183 的编号。此外，每位作者都被分配了一个“a_#”形式的 ID 标签，其中 # 从 0 开始，到元素总数减一为止。类似地，每个来源都被分配了一个“j_#”形式的 ID 标签，机构则被分配了一个“i_#”形式的标签。每个国家都被分配了一个“c_#”形式的标签。每位作者的关键词都被标记为“k_#”，每个扩展关键词都被标记为“p_#”。此外，如果文档中引用了参考文献，则会为其分配一个“r #”标签；唯一的例外_是该参考文献属于这 184 篇文档之一，此时保留其原有标签。这套标记系统有助于以更清晰简洁的方式识别和可视化不同的元数据元素。

以下部分将运用探索性数据分析、网络分析和人工智能技术，进一步分析元数据元素并生成见解。您可以通过bit.ly/3mcP2eb，访问这些步骤；该链接提供了一个用户友好的在线界面，方便您开展这些分析并更详细地探索数据。

除了使用提供的链接，用户还可以通过自己偏好的集成开发环境（IDE），使用以下命令直接安装该库：pip install pyBibX。安装后，即可在 Python 脚本中导入 pyBibX 软件包，从而使用其功能，并对自己的数据集进行文献计量分析。

4.1 pyBibX – 探索性数据分析

在本节中，我们将对合并数据集进行探索性数据分析。这是数据分析中的关键步骤，有助于我们了解数据的底层结构、模式和关系。此外，这一步骤还能全面概览合并数据集，并帮助我们识别可能为后续分析提供参考的关键特征和趋势。表04展示了初步统计报告。

表04 – 合并元数据报告

	结果
时间跨度	1984—2023年
国家总数	45
机构总数	231
来源总数	121
参考文献总数	7926
语言总数	1
--英语（文档数）	184
文档总数	184
--文章	184
每位作者的平均文献数	1.2
每个机构的平均文献数	2.64
每个来源的平均文献数	1.52
年均文献数	7.36
作者总数	495
作者关键词总数	587
作者关键词加总数 1434	
单作者文献总数	43
多作者文献总数	141
平均合作指数	3.21
最大 h 指数	21
总被引次数	5674
每位作者的平均被引次数	11.46
每个机构的平均被引次数	24.56
每篇文档的平均被引次数	30.84
每个来源的平均被引次数	46.89

表04列出了有关文档集的各项统计数据，该文档集涵盖1984—2023年。文档以英语撰写，来源于121种不同的期刊。涉及的国家共45个，参与文档产出的机构有231家。这些文档全部为文章，平均每位作者发表1.2篇文档、每家机构产出2.64篇文档、每个来源收录1.52篇文档。整个期间共有495位作者和587个作者关键词，以及1434个扩展作者关键词。在184篇文档中，43篇为单人署名，141篇为多人署名，平均合作指数为3.21，最高h指数为21。这些文档共获得5674次引文，平均每位作者获引11.46次、每家机构获引24.56次、每篇文档获引30.84次、每个来源获引46.89次。

从表2所呈现的信息中可以得出几个结论。以下是其中一些可能的结论：a）该文献集合包含184篇文档，规模相对较小，在涵盖的40年间平均每年仅有7.36篇文档。b）平均合作指数为3.21，表明这些文档是多位作者合作完成的，而非单人署名作品。c）每个来源的平均被引次数相对较高（46.89），这说明该集合中的文档参考文献较为丰富，引用的来源也多种多样。d）最高h指数为21，表明该集合中至少有一位作者被引次数很高。不过，每位作者的平均被引次数为11.46，这说明这种情况并非所有作者的普遍现象。e）每个机构的平均被引次数为24.56，这表明该集合中某些机构的占比可能高于其他机构。

虽然表04得出的结论提供了一些有价值的见解，但仍有若干方面值得进行更深入的分析。例如，表2提供了一些关于协作和引文指标的信息，但更详细的分析可以探讨文档集合中的协作性质与范围，以及作者、机构和来源之间的引文与参考文献模式。

继续进行探索性数据分析，一种进一步了解这些文档的方法是使用词云和N元语法进行文本分析。通过从文档摘要中生成词云和N元语法（n 是大于 1 的正整数），或许能够识别作者使用的常见主题、话题和术语。图01展示了这两种探索性分析的结果：一个词云和一个三元语法（包含 3 个词的 N 元语法）。

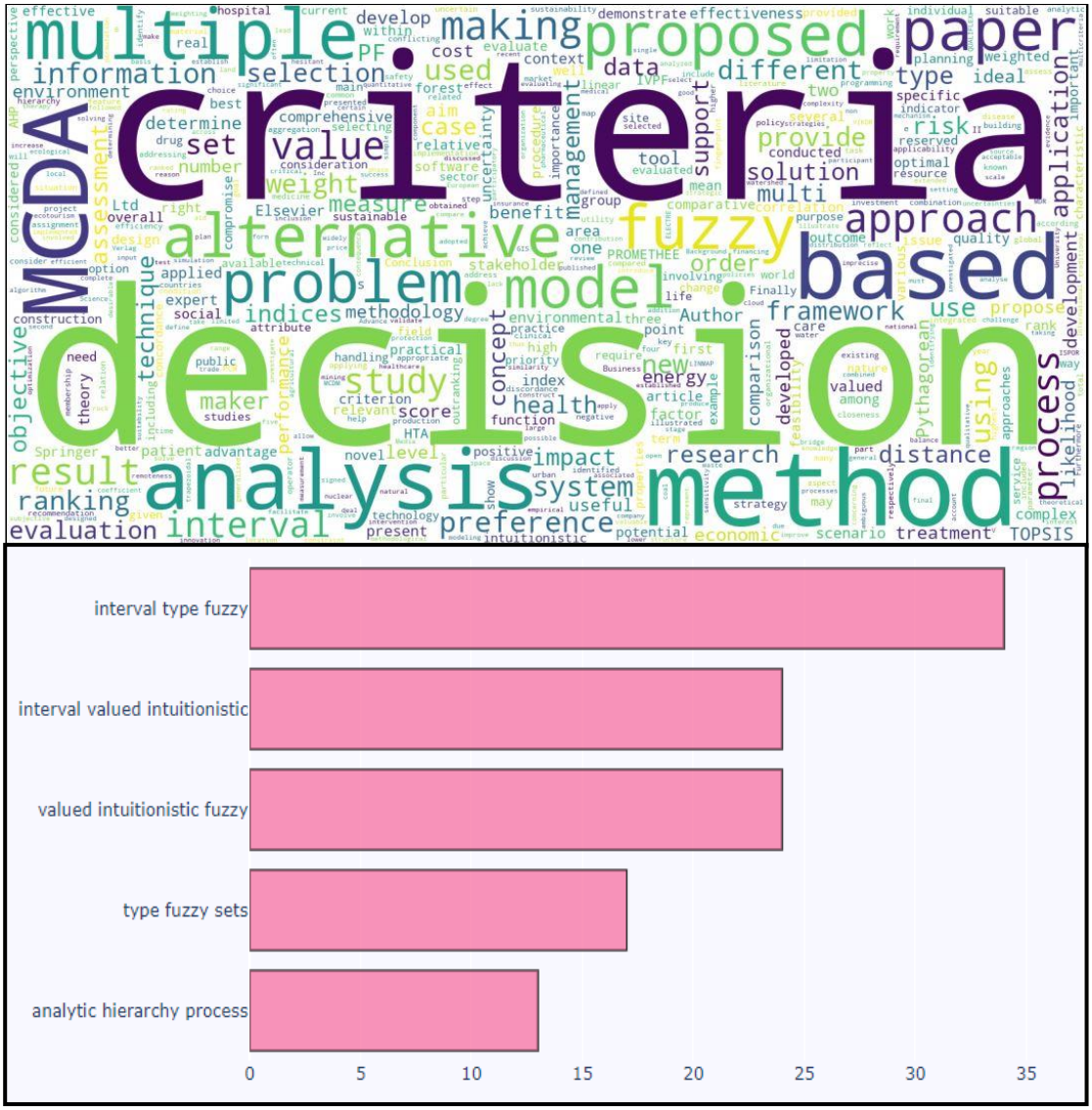


图01- 词云和三元语法

根据由文档摘要生成的词云（图01上方的图）中突出显示的词语，这个文献集似乎主要关注决策，尤其是多准则决策分析（MCDA）和模糊逻辑方法。词云中这些术语的突出程度表明，该文献集可能对关注决策过程和技术的研究人员及从业者有所裨益。进一步分析该文献集，可以探讨与多准则决策分析（MCDA）和模糊逻辑方法相关的具体应用、挑战与创新。

三元语法分析（图01下方的图）揭示了文档中讨论的具体概念和技术。不过，首先对文档摘要进行了规范化处理，移除了英语停用词以及以下该知识领域特有的自定义停用词：“analysis”（分析）、“criteria”（标准）、“decision”（决策）、“elsevier”、“making”（制定）、“mcda”、“multiple”（多个）、“reserved”（保留）和“rigths”（权利）。规范化后，出现频率最高的三元语法是“interval-type-fuzzy”，频率为34，表明这一概念是这批文档的核心主题。“interval-valued-intuitionistic”和“valued-intuitionistic-fuzzy”这两个三元语法的出现频率均为24，凸显了这些概念的重要性。

三元语法“type-fuzzy-sets”出现了17次，表明模糊集及相关技术也是值得关注的主题。有趣的是，三元语法“analytic-hierarchy-process”出现了13次，证实层次分析法（AHP）是广为人知的多准则决策方法（MAZUREK et al. 2020；MAZUREK et al. 2021；PEREIRA & COSTA，2015；FLORIANO et al. 2022）。在我们的库中，该图具有交互功能，将光标悬停在图上即可查看更多结果信息。

使用词云和N元语法等技术探索数据集后，探索性数据分析的下一步是将摘要转换为数值表示。184份文档的摘要可转换为词频-逆文档频率（TF-IDF）矩阵或嵌入。以这种方式表示文档后，可以应用UMAP或TSVD等降维技术，将维度降至两个，以便进行可视化。基于降维技术进行文档聚类，广泛用于分析文本；通过降低数据的维度，这些技术可以揭示文本数据的底层结构，使其更易于探索和解读（RAMKUMAR & POORNA，2016年）。最后一步是使用k均值算法（McQUEEN，1967年），根据距离度量将文档分为k个自然簇，并绘制结果图，每个簇以不同颜色突出显示。通过可视化这些簇，研究人员可以快速识别数据中的模式和关系，并利用这些信息更明智地决定分析的后续步骤，帮助研究人员发现新的见解和研究方向。

对于这种方法，通常更推荐将 UMAP 与嵌入结合使用，因为与 TSVD 相比，UMAP 是一种更先进的降维技术，也更适合保留数据的全局结构（JEON 等，2022年）。此外，UMAP 更擅长处理具有复杂结构的高维数据，这类数据在自然语言处理中十分常见。另一方面，由于 TSVD 和 TF-IDF 这两种技术速度较快，且能够处理大量文本数据，因此通常比 UMAP 和嵌入更受青睐（DESSI 等，2021年）。不过，需要注意的是，技术的选择最终应取决于所分析文本数据的具体特征。图02展示了使用 UMAP 与嵌入组合对 184 篇文档摘要进行的投影。

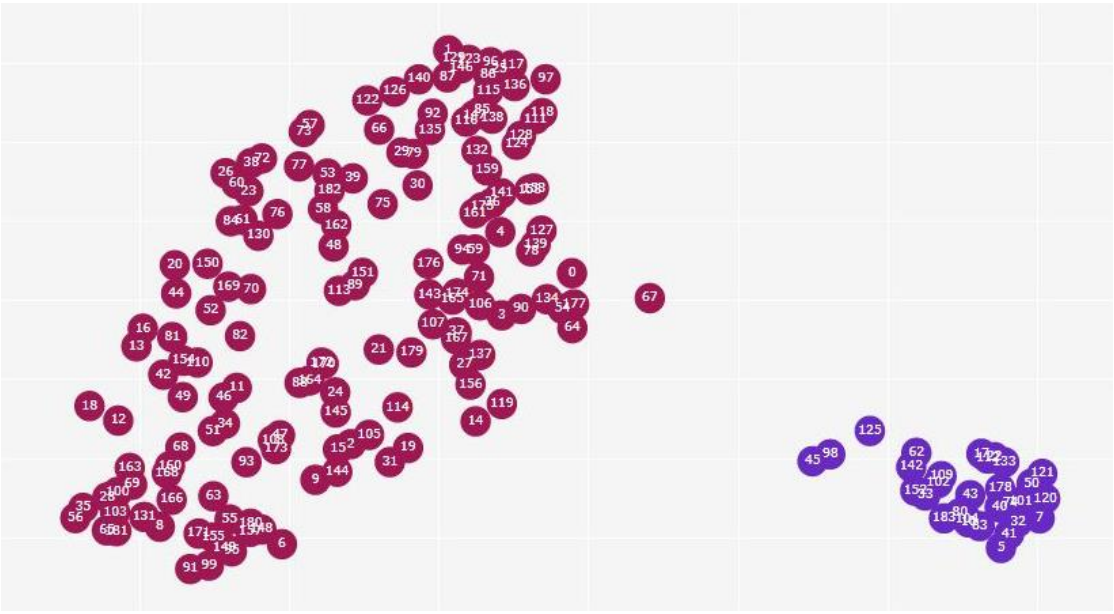


图02- 文档投影 - UMAP 和嵌入

在图02中，我们任意选择了 2 个簇，以直观呈现转换后的摘要。该图是一个交互式图表，将鼠标悬停在节点上即可查看其他信息，包括文档 ID、簇编号和完整的文献引文信息。通过投影，我们识别出一个与另一个簇明显不同的高度同质簇（紫色），这表明该簇中的文档可能共享一个

共同主题或话题。进一步研究该簇，或许能深入了解这一主题或话题，并有望发现新的研究方向。

识别主题或研究主题后，还可以使用各种技术进行进一步研究。一种方法是绘制多年间摘要中所使用关键词的演化图，这有助于了解该主题的演变情况以及正在出现的子主题。通过观察某些词语的出现频率和使用方式如何随时间变化，研究人员可以深入了解该领域关注重点的转移以及新主题的出现。图 03 展示了 2018年 至 2021年 关键词增强词的演化图。

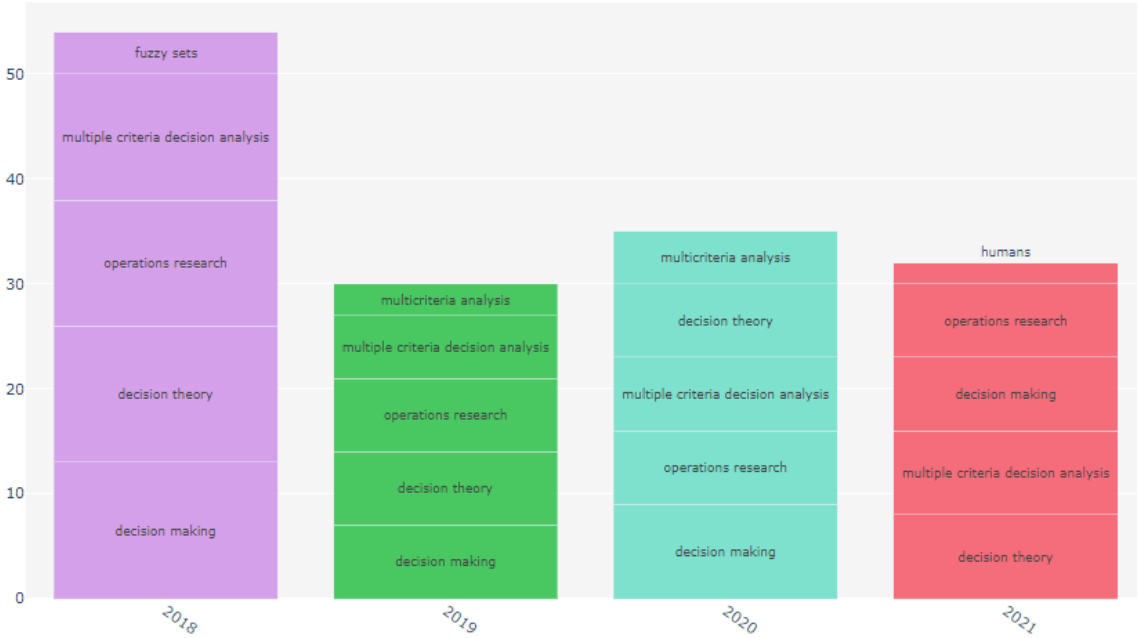


图 03- 演化图 2018年 – 2021年 关键词增强词

图 03 揭示了研究主题中的一些有趣趋势。在这四年中，最常出现的关键词是：“多准则决策分析”、“决策理论”、“运筹学”和“决策”。这些关键词表明，关于决策和分析的研究正在持续开展，并且十分活跃。尤其是“多准则决策分析”的高频出现，表明人们持续关注这种决策方法。另一个有趣的发现是，“人类”出现在 2021年 的关键词中，这表明人们对决策中的人为因素日益感兴趣。

这种技术能够直观呈现研究重点随时间的变化。图的交互特性便于探索特定关键词的出现频率，而时间维度也是一个关键因素。不过，需要注意的是，主题建模还可以挖掘更深层次的见解，并帮助我们更细致地理解潜在的主题和话题。

结合演化图，我们还可以使用树图，帮助理解全部 184 篇文档中的核心主题。通过分析这 184 篇文档中所有关键词增强词，我们可以迅速找出出现频率最高的关键词，这些关键词或许代表着重要的主题和议题。图 04 展示了排名前 15 的关键词增强词。

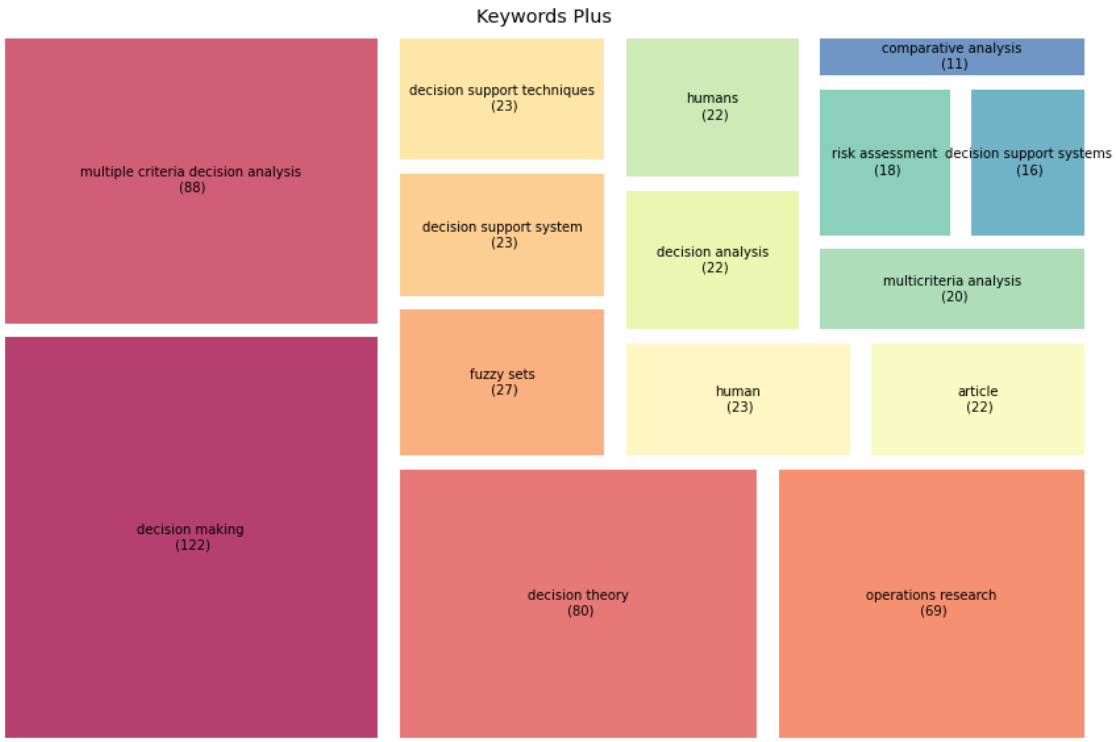


图 04—关键词增强词排名前 15 位树图

图 04 展示了 184 篇文档中关键词增强词的出现频率。最常见的是“决策”，出现 122 次；其次是“多准则决策分析”，出现 88 次；“决策理论”则出现了 80 次。“运筹学”和“模糊集”的出现频率分别为 69 次和 27 次。文档中“运筹学”这一关键词的高频出现表明，该学科与决策和多准则决策分析（MCDA）相关。 文档中“模糊集”这一关键词出现频率较低，表明与“运筹学”相比，这种方法不太常见。不过，这仍说明“模糊集”是这一领域的相关议题，值得进一步研究。同样值得注意的是，关键词“human”的出现频率为 23 次，“humans”为 22 次，这表明决策过程中关注以人为中心的方法。

如前所述，桑基图展示七类元素任意组合之间的文档流向：作者、国家、机构、来源、作者关键词、关键词增强词或语言。为了继续分析，我们可以展示作者与关键词增强词之间的流向，图 05 展示了排名前 10 的流向。

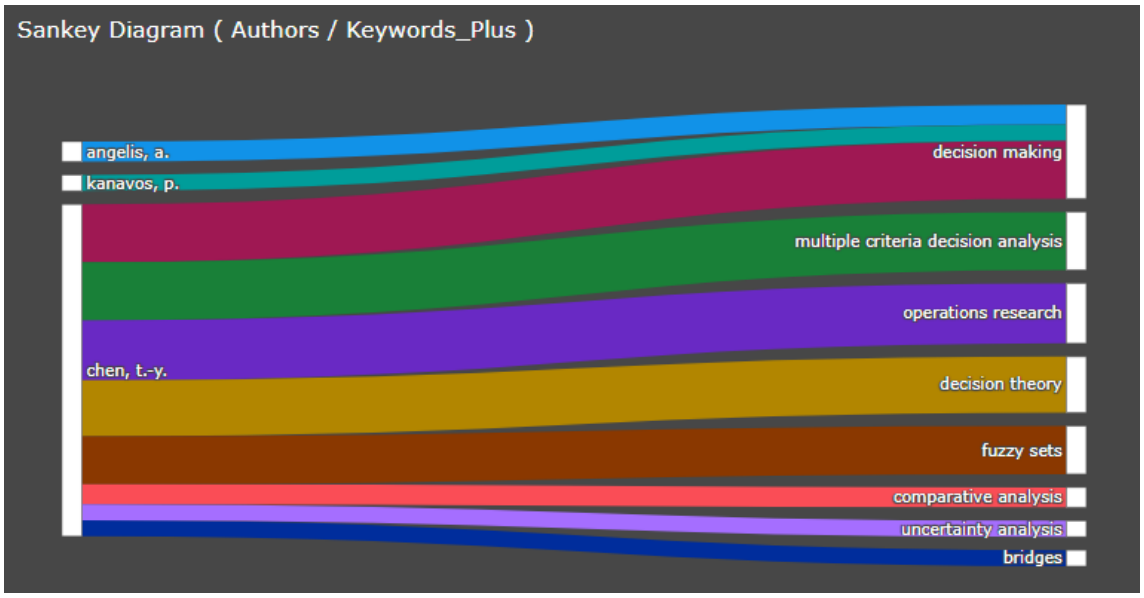


图 05- 桑基图 作者-关键词增强词

图 05 展示了作者与关键词增强词之间排名前 10 的流向：Angelis A. 和 Kanavos, P. 各有一条流向，Chen, T.-Y. 有八条。关键词增强词主要与树图（图 04）中识别出的主题一致，包括“决策”、“多准则决策分析”、“运筹学”、“决策理论”和“模糊集”。值得注意的是，Chen, T.-Y. 似乎是成果最多的作者，研究主题范围广泛。结果表明，作者之间可以围绕共同兴趣开展协作，从而为该领域作出更多贡献，并探索研究机会。

为了确认 Chen, T.-Y. 是一位高产作者，我们可以查看图 06 中的作者生产力图，该图展示了历年来生产力最高的 10 位作者。

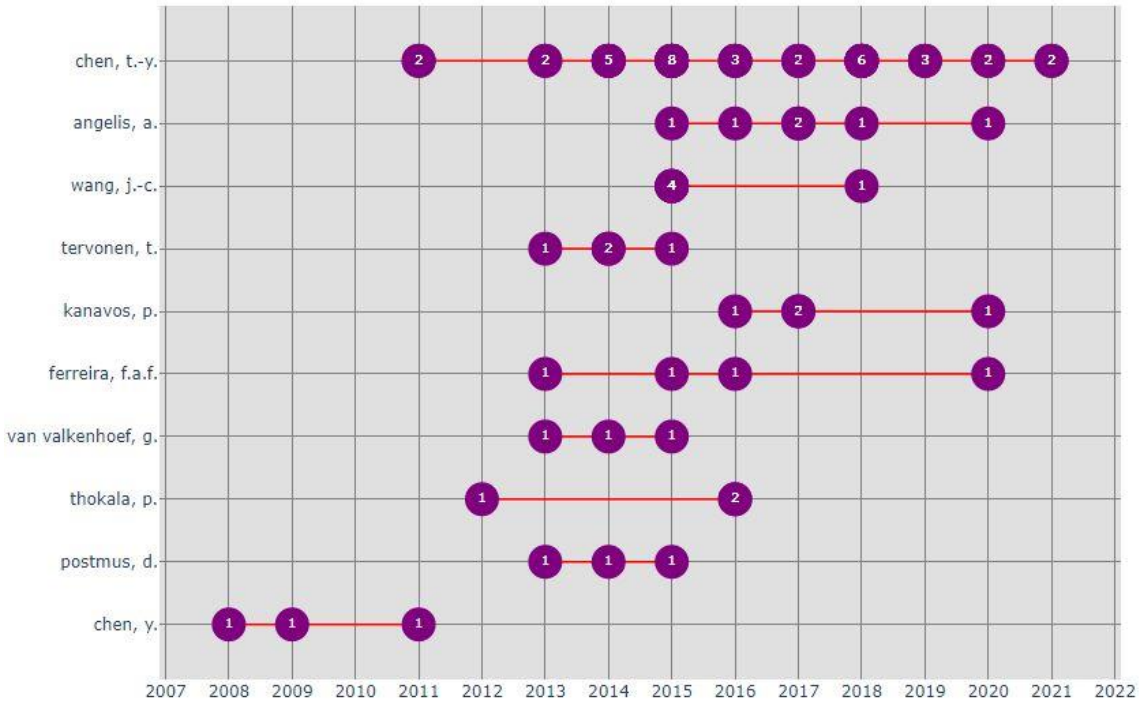


图 06- 作者生产力图-前 10 名

图 06 显示，Chen, T.-Y. 在 2011年 至 2022年 期间始终发表的文档最多。此外，我们还可以重点关注 Angelis, A 及其他生产力显著的作者。此图也具有交互功能，将鼠标悬停在节点上即可显示文档 ID，提供有关各位作者出版物的更多信息。通过考察最杰出作者的出版物，研究人员可以更好地了解该领域当前的研究状况，并找出需要进一步研究的方向。

最后，我们可以通过条形图核实有关元数据的一些额外统计信息，从而结束探索性数据分析。可显示的选项包括：每年文档数、每年引用数、每年过往引用数、洛特卡定律（LOTKA，1926年）、每篇文档的来源数、每次引用的来源数、每篇文档的作者数、每次引用的作者数、每个 h 指数的作者数、布拉德福定律、每篇文档的机构数、每次引用的机构数、每篇文档涉及的国家数、每次引用涉及的国家数、每篇文档的语言、每篇文档的扩展关键词，以及每篇文档的作者关键词。继续分析，我们可以核实文档产出是否逐年下降，因此图 07 展示了每年文档数图。

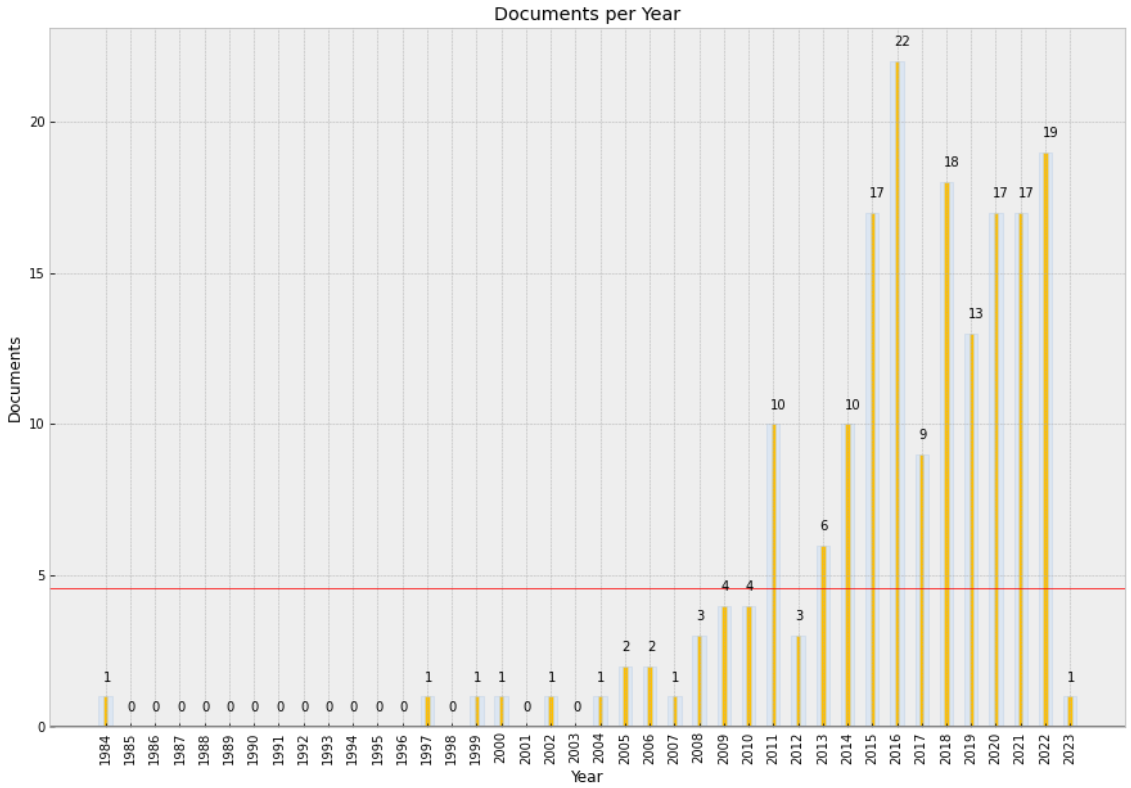


图 07—条形图—每年文档数

图 07 显示，该领域发表的文档数量呈上升趋势，表明人们对这项研究的兴趣日益浓厚。令人鼓舞的是，在整个期间内，每年发表的文章平均数量接近 5 篇，这表明该领域一直保持着稳定的关注度。此外，考虑到过去五年（2018—2022年），每年发表的文档平均数量已增至近 17 篇，说明人们对这一领域的兴趣更加浓厚。这些结果表明，研究人员有许多机会为该领域作出贡献并开展协作。

4.2 pyBibX – 网络分析

完成探索性数据分析后，我们可以继续进行网络分析，该分析侧重于科学出版物、作者和期刊之间的关系。在网络分析中，出版物、作者和期刊被表示为节点，它们之间的关系则被表示为边。网络分析可以识别作者或机构之间的合作模式、描绘某个领域中的思想脉络，或找出特定研究领域中的重要出版物或期刊（DING，2011年）。

邻接分析研究实体之间的关系或连接，通常以网络或图的形式表示。在引文分析（邻接分析的一种类型）中，文档之间的连接由

节点之间的连边表示。每个节点代表一篇文档，连边则代表一篇文档对另一篇文档的引文。在我们的方法中，每篇文档以蓝色节点表示，其参考文献以红色节点表示。通过可视化这些节点及其连接，我们可以了解文档之间的关联，以及哪些参考文献在多篇文档中被频繁引用。图 08 展示了文档子集的引文分析，以及至少被引用四次的参考文献。

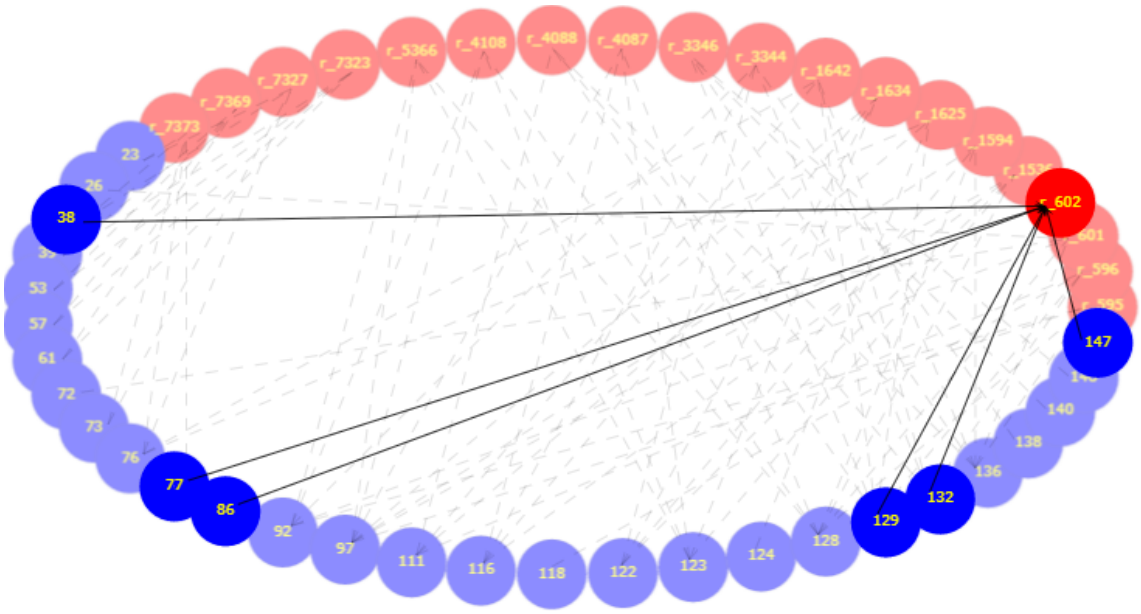


图 08- 引文分析

图 08 展示了上述文档和参考文献子集，并突出显示了参考文献“r_602”（ATANASSOV，1986年）。该文献被文档“38”（CHEN，2020年）、“77”（CHEN，2018年）、“86”（ZHANG等人，2017年）、“129”（CHEN，2014a年）、“132”（CHEN，2014b年）和“147”（CHEN，2011年）引用。该子集包含 19 条参考文献（红色节点）和 28 篇文档（蓝色节点），与数据集中的 184 篇文档总数相比，这个数量可能显得较少。不过，值得注意的是，该子集是根据至少被引用四次这一标准专门选出的，说明这些文献在该研究领域具有较高的重要性和相关性。突出显示的参考文献展示了 Atanassov 关于直觉模糊集的研究，该研究被 Chen, T.-Y. 引用五次，并被 Zhang等人引用一次。这表明 Chen, T.-Y. 可能对这一主题尤为感兴趣，因为多年来他多次引用了这篇参考文献。

引文历史分析是了解特定研究领域演变的有力工具。通过分析这184篇文档之间的引文，可以识别出重要贡献、关键作者和新兴趋势。这类分析有助于研究人员发现文献中的空白、潜在合作者以及研究的新方向。此外，它还有助于追踪特定出版物随时间推移所产生的影响，以及该出版物对后续研究的影响。图09展示了引文历史分析，其中突出显示了文档“97”。

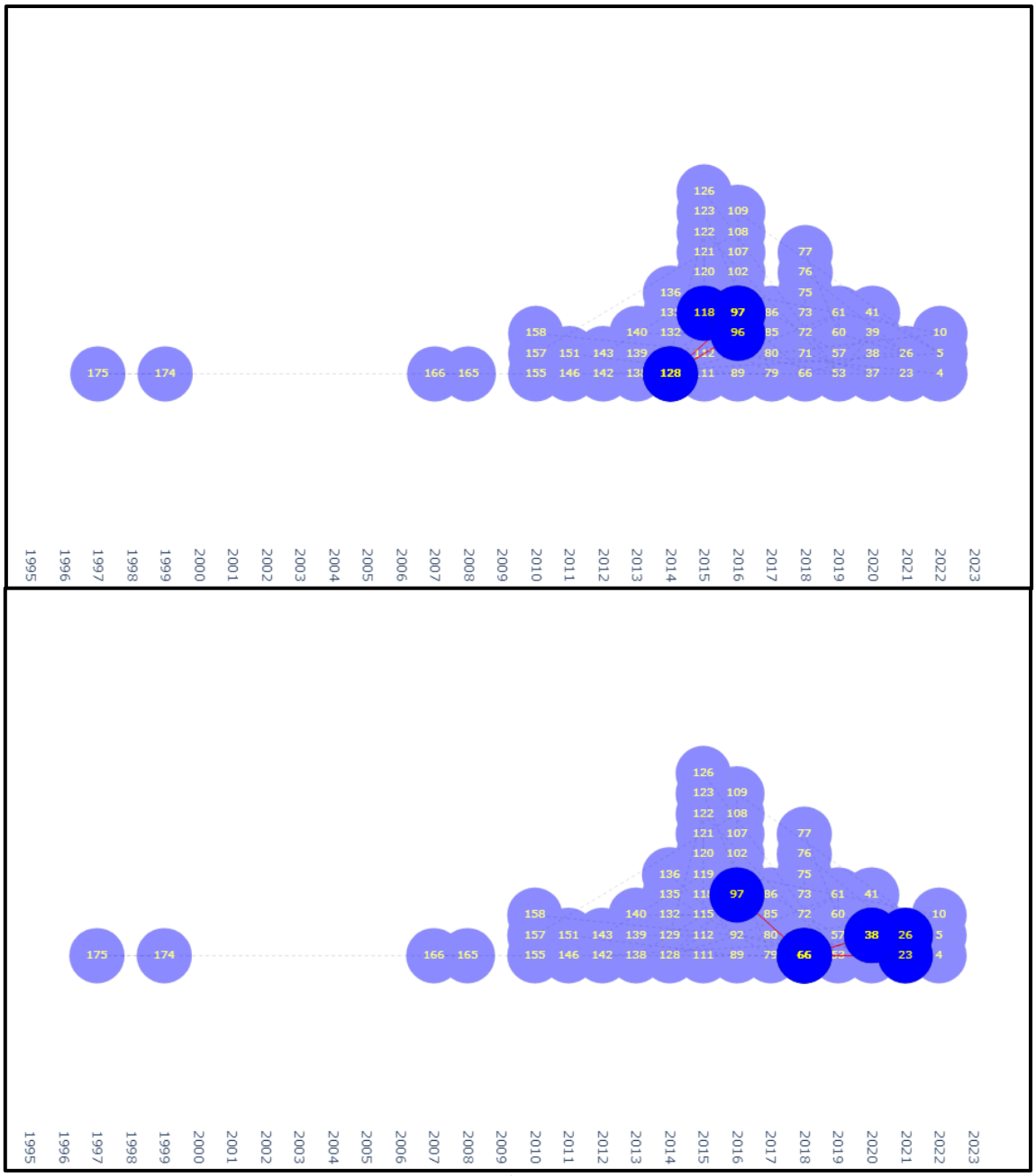


图 09- 引文历史分析

图 09 展示了一个引人入胜的文档网络，这些文档均被标记为“97”的中心文档引用。上方的图揭示了由“97”引发、延伸至早期文档的连锁反应。根据该网络，“97”引用了“118”和“96”，而“97”和“96”均引用了“128”。下方的图展示了同一连锁反应，但这次是朝着后续文档延伸。标记为“66”的文档被发现引用了“97”。文档“38”和“23”引用了“66”，最后，“26”引用了“38”，由此构成完整的网络。这种对引文网络的全面可视化，为理解不同文档和作者之间的关系提供了宝贵见解，使研究人员能够对特定领域随时间的演变得出许多结论。通过考察单篇作品和作者的引文模式，研究人员可以识别出推动某一领域发展的协作关系和所属机构。在此示例中，所有文档均为 Chen, T.-Y. 发表的作品。

根据我们的探索性数据分析，以发表成果最多的研究人员 Chen, T.-Y. 为中心开展合作分析，将会很有意思。根据 Castañer 和 Oliveira（2020年）的说法，合作分析是一种用于研究为实现共同目标而协作的个人、群体或组织之间互动与关系的方法论。合作分析的结果有助于了解合作网络的结构与运作方式，以及如何利用这些网络来促进

协作、知识共享与创新。图 10 展示了研究人员 Chen, T.-Y. 合作网络的扩展情况。

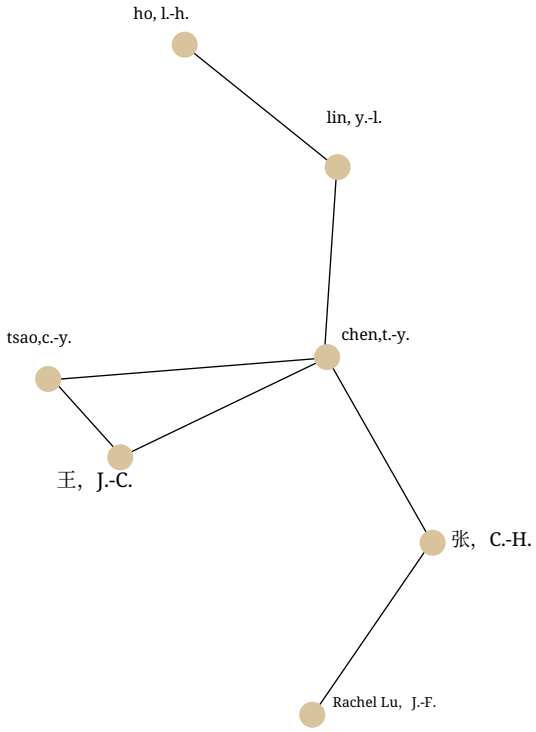


图 10—协作网络分析

图 10 显示，Chen, T.-Y. 仅与另外四位研究人员合作过，包括 Lin, Y.-L.、Chang, C.-H.、Tsao, C.-Y. 和 Wang, J.-C. 此外，Tsao, C.-Y. 与 Wang, J.-C. 合作过，Lin, Y.-L. 与 Ho, L.-H. 合作过，而 Chang, C.-H. 与 Rachel Lu, J.-F. 合作过。在现有的 184 份文档中，Chen, T.-Y. 撰写了 35 份，因此显然是该领域产出最多、影响力最大的研究人员。然而，他仅与另外四位研究人员合作这一事实表明，如果条件允许，该领域仍有开展更多协作和跨学科研究的空间。

我们可以采用类似的方法，将关注点转向文档，并考察它们如何彼此“协作”。由于文档无法以传统意义上的方式协作，因此我们需要进行相似性分析。相似性分析是探索不同文档之间关系的一种重要方法，可通过多种方法开展。一种常用方法是书目耦合和共被引分析，即分析不同文档中引用的参考文献，以识别它们之间的模式和关系（BOYACK and KLAVANS，2010）。在本例中，我们将使用共被引分析来识别具有共同参考文献的文档。具体而言，我们将查找至少有十条共同参考文献的文档，从而识别主题内容及所依据来源都高度相关的文档。图11展示了对184份文档进行的共被引分析。

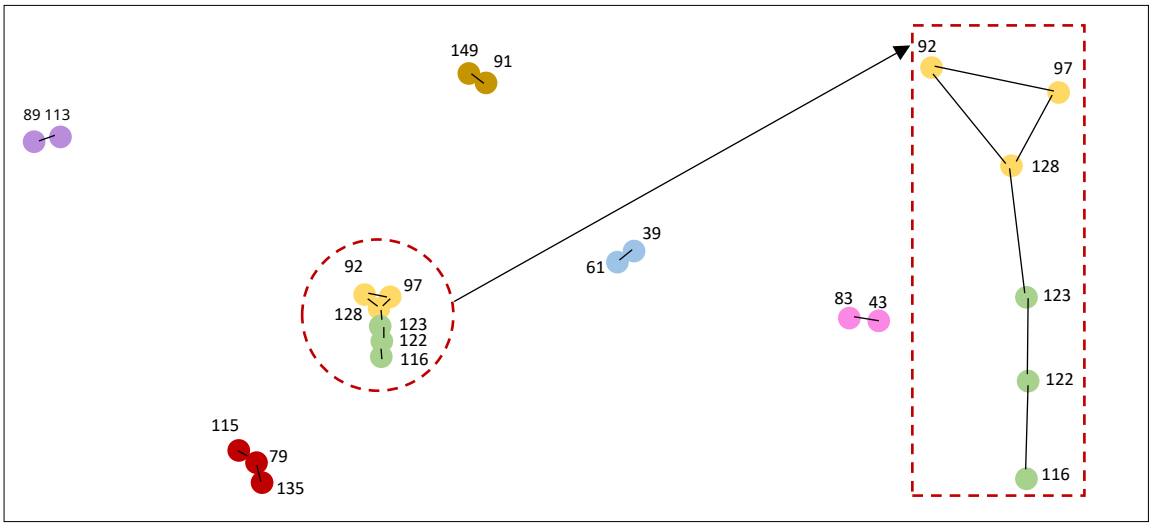


图11- 相似性分析 – 共被引分析

在图11中，我们可以看到七个不同的簇，分别以紫色、红色、黄色、绿色、棕色、蓝色和粉色标示。放大黄色和绿色簇后，可以看到它们之间的连线，这构成了两个簇之间的重要联系。紫色簇包含文档“89”和“113”；棕色簇包含文档“149”和“91”；红色簇包含文档“115”、“79”和“135”；黄色簇包含文档“92”、“97”和“128”；绿色簇包含文档“123”、“122”和“116”；蓝色簇包含文档“61”和“39”；最后，粉色簇包含文档“83”和“43”。 有趣的是，研究发现，Chen, T.-Y. 在黄色、绿色、蓝色和红色簇中均有文档，表明该作者的研究主题覆盖范围广泛。

分析黄色和绿色聚类可知，配对“92-97”有14篇共同参考文献，“92-128”有10篇，“97-128”有20篇。黄色与绿色聚类之间的连线“122-123”有10篇共同参考文献。在绿色聚类中，“123-122”有26篇共同参考文献，“122-116”有13篇。根据分析可以得出结论：黄色和绿色聚类中的文档具有显著的相似性，这一点可从它们之间大量共享的参考文献中得到印证。两个聚类之间的连线（文档对“122-123”）也显示出中等程度的相似性。尤其是绿色聚类中的文档具有较高的相似性，其中“123-122”和“122-116”这两组配对拥有许多共同参考文献。这些发现表明，黄色和绿色聚类中的文档可能具有共同的研究主题。

科学研究依托合作蓬勃发展，而分析合作网络能够深入揭示研究人员和机构如何携手推进知识发展。为全面了解科学知识的产出与共享方式，可以开展世界地图合作分析，探究不同国家的研究人员和机构如何合作开展研究项目。此类分析能够揭示国际合作中的模式与趋势，识别关键参与者和机构，并为资助决策和政策制定提供参考。图12展示了世界地图合作分析。

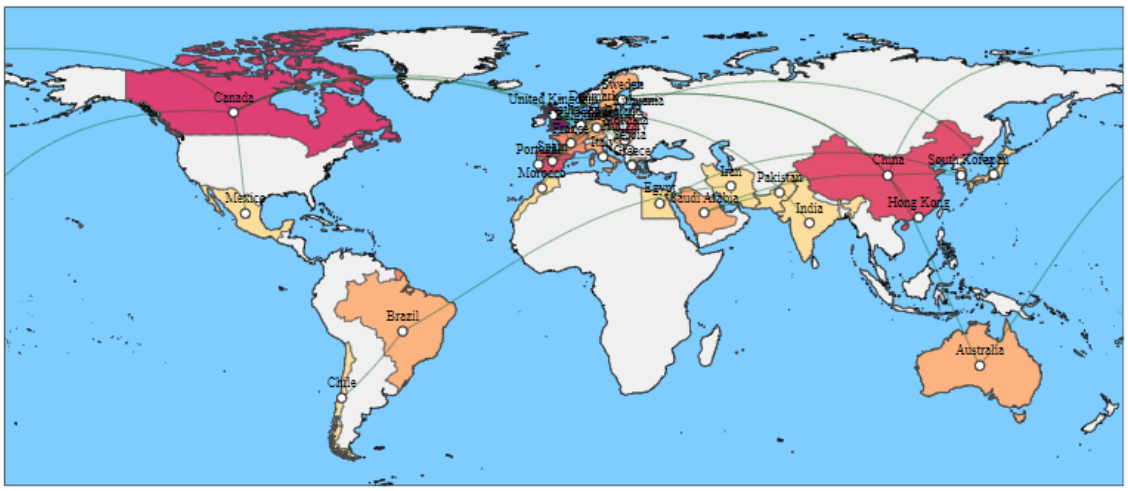


图12-世界地图合作分析

图12展示了世界地图合作分析的结果，凸显了合作在科学研究中的重要性，并揭示了国际合作中一些有趣的模式。图中显示，欧洲研究人员往往更频繁地与欧洲同行合作，这表明欧洲内部形成了稳固的合作网络。加拿大和中国成为重要的合作枢纽，说明这些国家在促进全球科学知识产出方面发挥着关键作用。另一方面，分析还发现，巴西、智利和澳大利亚等国家的合作网络较弱，这可能表明有必要加大力度，推动这些地区的合作。世界地图合作分析为了解国际科学合作的模式和动态提供了宝贵见解，可为未来促进跨境合作与知识交流的工作提供参考。

4.3 pyBibX – 人工智能分析

完成探索性数据分析和网络分析后，下一步是使用人工智能工具，深入了解文档集合。具体而言，基于深度学习技术的嵌入、主题建模和文本摘要，有助于揭示数据中隐藏的主题和模式。通过自动识别并总结关键主题和趋势，这些工具能够以更高效、更准确的方式，从大量文本数据中提取含义。

我们的研究将 UMAP 与嵌入相结合，将摘要投影到低维空间，并对其关系进行可视化。嵌入采用 Reimers 和 Gurevych（2019年）开发的 Sentence-BERT 深度学习模型生成。除了可视化摘要之间的关系，嵌入还可作为其他深度学习技术的输入。将这些嵌入用作输入后，深度学习模型能够更好地捕捉文本的语义含义和上下文。这些表示可以改进关键主题识别或生成简明摘要等任务的结果，展现了嵌入在各种自然语言处理任务中的多功能性和实用价值。

主题建模是一种能够自动检测文档集内隐藏的主要主题和模式的方法。通过分析文本中的词语和短语模式，可以识别出经常共同出现的词语群组，并将其解读为主题，代表文档集合中潜在的概念或思想。主题建模的主要目的是提供一种方法，以便从大量文本数据中识别并理解最重要的主题。表05展示了使用 Grootendorst（2022年）开发的 BERTopic 深度学习模型进行主题建模的结果。

表05 – 主题建模

	相关词语	主题实验室 **el**	中心文档
主题 0 (153)	类型、选择、排序、数据，集合、偏好、过程，信息、区间、模糊	决策与偏好分析	“96”（CHEN，2016年）
主题1 (31)	益处、影响、治疗，评估、患者、风险，框架、护理、卫生技术评估、健康	卫生技术评估与患者护理	“74”（ANGELIS，2018年）

根据表05所呈现的结果，可以得出结论：主题建模已成功识别出这184份文档中的两个主要主题。主题0最为突出，涵盖153份文档；而主题1仅涵盖31份文档。与主题0相关的词语表明，其重点是“决策与偏好分析”；有助于描述该主题的一组词语也支持这一主题标签。Chen（2016年）的中心文档“96”主要体现了该主题。另一方面，从相关词语来看，主题1似乎侧重于“卫生技术评估与患者护理”。其中心文档是 Angelis（2018年）的“74”。这些发现表明，该文档集合以决策与偏好分析研究为主，另有较小一部分研究与卫生技术评估和患者护理相关。此外，图02所呈现的证据也支持这些发现。

已识别的中心文档“96”（CHEN，2016年）和“74”（ANGELIS，2018年），可以使用抽象式摘要和抽取式摘要技术进行概括。抽象式摘要是指创建新的摘要，传达原文中的核心信息。相比之下，抽取式摘要则是从原文中选取最重要的句子或短语，形成精简版本。这两种技术的主要区别在于，抽象式摘要需要生成能够准确传达原文含义的新文本；而抽取式摘要则是从原文中选取并精简文本。抽象式摘要更具挑战性，但可能更完整地传达原文含义；抽取式摘要则更直接，但可能无法完整传达原文含义（YADAV 等，2022年）。

这两种摘要模型都采用了先进技术；具体而言，抽取式摘要是使用 Derek（2019年）构建的 BERT 模型完成的。尽管 BERT 在众多自然语言处理任务中确实取得了最先进的成果，但需要认识到，要在特定领域的文档中实现最佳性能，可能需要进行微调。使用领域特定语料库对 BERT 进行微调，能够更好地捕捉专业主题的细微差别、术语和语境。这种定制方法可确保 BERT 更准确、实用地应用于领域特定任务，从而提升性能，并加深对特定领域独有特征的理解。

另一方面，抽象式摘要通过 PEGASUS（一种由 Zhang等人（2019年）开发的先进方法）和 ChatGPT（OpenAI，2023年）实现；后者是当前大语言模型（LLM）领域的最先进技术，且并非总能免费使用。表 06 展示了这些技术的结果，输入均为摘要。文章的原始摘要可通过其唯一的数字对象标识符，在相应科学期刊的网站上轻松查阅。

表 06 – 文本摘要

中心文档	抽象式摘要 - PEGASUS
“96”（CHEN，2016年）	作者隶属于维尔纽斯格迪米纳斯技术大学（VGTU）[sic]
“74”（ANGELIS，2018年）	在卫生技术评估中使用多准则决策分析已被 视为一种前瞻性方法论，可协助决策者评估 卫生技术评估中新药的益处 [sic]
中心文档	抽象式摘要 - ChatGPT
“96” (CHEN, 2016年)	区间值直觉模糊集（IVIFSs）提供了一种处理不确定 和模糊属性的方法。通过包含比较 可能性对 IVIFSs 进行比较，从而得到一个基于包含关系的指标。该指标用于构建 线性规划模型（用于信息一致的情况）和集成非线性 规划模型（用于信息不一致的情况），以估算准则权重并 在不完整的偏好结构下确定备选方案的最优排序。 所提出的方法如 [sic] 所示。
“74”（ANGELIS，2018年）	多准则决策分析（MCDA）是一种已被认定为 一种有望协助决策者评估新药益处的方法 在卫生技术评估（HTA）中。然而，相关的实证证据有限。 从真实世界应用中测试此方法论。为了测试多准则决策分析 在现实世界环境中的方法论中，四位专家参加了一次决策会议 他们使用 Advance 价值框架评估三种不同治疗的益处 转移性去势抵抗性疾病的治疗 [sic]
中心文档	抽取式摘要 - BERT
“96”（CHEN，2016年）	摘要：区间值直觉模糊集理论提供了一种直观且 可行的方法，用于处理不确定和模糊属性。许多有用的模型和 方法已被开发出来，用于在区间值直觉模糊 环境中进行多准则决策分析。[sic]
“74”（ANGELIS，2018年）	背景。多准则决策分析（多准则决策分析）已被认定为一种 用于协助决策者评估新 药在卫生技术评估（卫生技术评估）中的益处；然而，来自真实世界应用的实证证据有限 。测量。参与者 s’ value preferen证据我们 re 通过标准选择、选项评分、标准权重分配及其 聚合而获得。结论。[sic]

表 06 展示了中心文档“96”（Chen，2016年）和“74”（Angelis，2018年）的抽象式摘要和抽取式摘要。将原始摘要与生成的摘要进行比较后可以看出，PEGASUS 和 BERT 生成的摘要遗漏了原始摘要中的重要信息。ChatGPT 生成的抽象式摘要则对原始摘要进行了更连贯、简洁的呈现。

关于 CHEN 的研究，PEGASUS 的抽象式摘要仅提供了作者所属机构的信息，而这与文档内容无关。另一方面，ChatGPT 的抽象式摘要抓住了原始摘要的核心，讨论了区间值直觉模糊集、包含比较可能性，以及为估算准则权重和最优排序而开发的线性与非线性规划模型。抽取式摘要强调了区间值直觉模糊集理论及其在多准则决策分析中的用途。然而，它未能全面概述所提出模型的方法论或应用。

关于 ANGELIS 的研究，PEGASUS 的抽象式摘要仅笼统提及在卫生技术评估中使用多准则决策分析。ChatGPT 的抽象式摘要内容翔实，涵盖了研究的关键方面，例如多准则决策分析在卫生技术评估中的潜力、Advance Value Framework 的真实世界测试，以及对三种转移性去势抵抗性前列腺癌治疗方案的评估。同时，抽取式摘要涉及研究背景、测量和结论，并指出多准则决策分析是卫生技术评估中一种颇具前景的方法论。不过，它缺少有关 Advance Value Framework 和所评估具体治疗方案的详细信息。

尽管 PEGASUS 和 BERT 摘要技术尚未达到与 ChatGPT 同等的熟练程度，但必须强调，在使用 ChatGPT 时存在显著差异。

应用，因为它需要付费 API 密钥，或订阅才能访问更高级的模型。同时
目前，PEGASUS 和 BERT 仍可免费使用，这使它们能够惠及更广泛的用户和研究人员。此外，随着这些技术不断演进，它们的性能最终可能达到甚至超越 ChatGPT。随着人工智能技术迅速发展，这些免费提供的摘要方法拥有巨大的开发和改进潜力，最终将成为那些无力或不愿投资使用 ChatGPT 等付费替代方案的人们的宝贵资源。

在最后这一步中，我们展示了 ChatGPT 与我们的库如何无缝集成。这得益于专门设计的函数，它们简化了对各类文献计量结果的检查与解读，包括探索性数据分析报告、词云、N元语法、演化图、桑基图、作者生产力图、条形图、引文分析、合作分析、相似性分析和世界地图合作分析。人类的求知欲与 ChatGPT 相结合，有望加深理解，并为传统的文献计量学方法带来新视角。

以图 07（按年份统计文档数量的条形图）为例，生成条形图后，研究人员可以提出各种有助于更好地解读结果的问题。我们将依次提出一系列不同的问题，以演示这一方法。

查询 1： 请分析所提供的信息

回答: 数据显示，1984年至2022年间，文档数量总体呈上升趋势。1984年至1995年间，数量几乎没有波动，除了1984年和1997年外，各年的文档数量几乎为 0。1996年至2021年间，文档数量缓慢但稳步增长，这一趋势在最近十年（2010—2020年）尤为明显。文档数量在2016年达到最高，2020年位居第二。2017年文档数量降至 9，随后在2018年和2019年分别回升至 18 和 13。2021年，文档数量再次增至 17。看来未来几年文档数量预计将维持在这一水平。 [sic]

问题 2： 2004年之前和之后发生了什么?

回答： 2004年之前，每年发布的文档数量相当少，大多数年份都没有新文档发布。2004年发布了一份新文档。2004年之后，每年文档数稳步增加，2005年和2006年各发布了两份文档，2008年发布了三份。文档发布数量继续稳步增长，并在2016年达到峰值，当年发布了22份文档。此后，文档发布数量略有下降，但与2004年之前相比仍保持在相对较高的水平。 [sic]

问题 3： 请列出提高人们对这一主题的研究兴趣的方法。

回答： 1. 分析数据中的趋势，确定推动或阻碍研究兴趣增长的因素。2. 寻找能够帮助拓展研究范围的潜在合作者。3. 利用社交媒体和其他数字平台宣传研究，吸引更多关注。4.制定一份全面且引人入胜的研究计划书，阐明项目的目标与宗旨。*5.* 制定外展策略，与公众互动并提高公众对研究的认知。6. 利用现有研究和数据为研究提供参考，并补充更多背景信息。7. 制定全面的数据收集与分析计划，以获取必要信息。8. 以直观且全面的形式呈现数据，更好地传达研究发现。9. 在知名期刊上发表研究论文，提高研究的可见度。10. 参加相关活动和会议，与其他研究人员交流并讨论想法。 [sic]

向 ChatGPT 提出的三个查询，重点是从数据中提取有关每年发表的文档数量的见解。第一个查询旨在更全面地了解相关信息。相比之下，第二个查询则具体要求提供 2004年之前和之后各年份的详细信息。最后，第三个查询询问了一系列方法，以提升人们对所研究主题的兴趣。ChatGPT 生成的回答质量良好，清晰简洁地阐述了数据中的见解，并为提升研究兴趣提供了有价值的建议。不过，回答仍有改进空间，尤其是前两个查询之间的衔接：第二个回答主要

重复了第一个回答的内容。一个可行的方法是设计更复杂的查询，鼓励 ChatGPT 更深入地分析数据并探索新颖的方面，从而生成更多样、更具信息量的回答。

值得注意的是，我们模拟了 chatGPT 如何展现出激发人类好奇心的非凡能力，这一点从三个查询中问题逐步深入的特点便可见一斑。第一个查询旨在概括性地了解数据，由此引发了对某个特定年份的明确兴趣，进而提出第二个查询。随着好奇心不断增强，第三个查询则转向寻求该主题的前瞻性指导。在这一循环过程中，研究者还可以继续深入探索更丰富的层面；chatGPT 巧妙地回应了研究者的提问，展现出培养求知欲、助力深入探索底层数据的能力。

4.4 pyBibX – 讨论

结合其研究目标来看，pyBibX 的理论应用涵盖多个维度，有助于深化对文献计量分析的理解，并凸显其在塑造科学领域方面的重要作用。通过融入先进的人工智能技术，pyBibX 提升了文献计量分析的方法论水平，为创新方法开辟道路。换言之，pyBibX 运用人工智能技术和人工智能辅助对话工具，增强了文献计量学的理论基础，并促成一种前所未有的新型协同效应。同时，对视觉美感的重视加深了对数据的理解，并推动了可视化理论的发展，也为科学研究中的有效交流确立了最佳实践。

值得注意的是，pyBibX 将先进的人工智能功能集成到代码中，例如主题建模、嵌入向量、文本摘要以及通用自然语言处理任务，从而提升了探索性数据分析和网络分析的结果。这些人工智能功能的融合提高了文献计量分析的准确性和效率，揭示了传统方法可能忽略的独特模式和趋势。在实践中，pyBibX 通过减少此类任务通常所需的人工投入和时间，简化了文献计量分析流程。效率的提升使科学家能够更迅速地响应各自领域的新发现和新进展。

pyBibX 的一项显著特点是能够促进研究人员与 ChatGPT 之间的互动，使他们能够提出问题，并根据分析结果获得全面的解答。这种互动方式加深了对数据的理解，并加快了隐性知识的发现，最终推动科学研究的发展。凭借创新功能，pyBibX 为研究人员提供了一款功能强大且富有洞察力的文献计量分析工具。

随着科学界不断发展壮大并积累海量数据，对高效、精细的文献计量分析工具的需求日益重要。pyBibX 通过提供一套由人工智能驱动的先进解决方案来满足这一需求，使研究人员能够灵活地探索复杂数据集。它开创性地融合了先进的人工智能能力，为研究人员提供了保持学科前沿地位的手段，帮助他们揭示隐藏的联系，并为集体知识库添砖加瓦。

5. 结论

pyBibX 库是学术界的一项重要资源，在分析科学数据、识别模式与趋势，以及生成促进协作的深刻见解可视化方面具有不可估量的优势。该库的效能在于能够从复杂数据中提取有意义的见解、自动执行特定任务，并确保数据分析的准确性与一致性，从而降低人为错误的风险。我们的文献计量学研究采用多方面方法，通过探索性数据分析、网络分析和人工智能分析，结合嵌入、主题建模和文本摘要等深度学习方法，并使用 BERT、Sentence-BERT、BERTopic、Pegasus 和 ChatGPT 等前沿工具。各项技术的成果能为研究人员节省大量时间和资源，使他们能够专注于假设构建、实验设计和论文撰写等更高层次的任务。简而言之，pyBibX 是学术界的一种工具，可丰富书目信息的分析与综合。

参考文献

AHMED, I.; SUTTON, A.J.; RILEY, R.D.（2012年）。出版偏差、选择偏差和
使用个体参与者数据进行荟萃分析时不可用的数据：数据库调查。BMJ。第344卷：d7762。数字对象标识符：10.1136/bmj.d7762

ANGELIS, A.（2018年）。使用新药的益处进行卫生技术评估
多准则决策分析：瑞典转移性前列腺癌患者使用牙科与药品福利机构（TLV）的案例研究。MDM Policy and Practice。第3卷，第2期，数字对象标识符：10.1177/2381468318796218。

ARIA M.；CUCCURULLO, C.（2017年）。Bibliometrix：用于全面科学图谱分析的 R 工具。信息计量学杂志。第11卷，第4期，第959-975页。数字对象标识符：10.1016/j.joi.2017.08.007

ATANASSOV, K.T.（1986年）。直觉模糊集。载于：直觉模糊集。模糊性与软计算研究。第35卷。Physica，海德堡。数字对象标识符：10.1007/978-3-7908-1870-3_1

BELLIS, N.D.（2009年）。文献计量学与引文分析：从科学引文索引到网络计量学。Scarecrow Press。国际标准书号 978-0-8108-6714-7。

BIRCAN, T.; SALAH, A.A.A.（2022年）。社会科学中人工智能技术应用的文献计量分析。《数学》，第10卷，第4398期。数字对象标识符：10.3390/math10234398

BÖRNER, K.; CHEN, C.; BOYACK, K.W. (2003). 知识领域可视化。载于 Cronin, Blaise】【。（编者），《信息科学与技术年度评论》第37卷，第1期，第179–255页。数字对象标识符：10.1002/aris.1440370106

BOURNE, C.P.; HAHN, T.B.（2003）。《在线信息服务史，1963–1976》。麻省理工学院出版社。国际标准书号 978-0-262-26175-3。

BOYACK, K.W.; KLAVANS, R. (2010). 共被引分析、书目耦合与直接
引文：哪种引文方法最准确地代表研究前沿？美国信息科学与技术学会会刊。第61卷，第2389-2404页。数字对象标识符：10.1002/asi.21419

BRADFORD, S.C.（1934年）。特定主题的信息来源。《工程：画报》
学期刊，第 137 卷，第 85-86 页。数字对象标识符：10.1177/016555158501000

CASTAÑER, X., & OLIVEIRA, N.（2020年）。组织间的协作、协调与合作：通过系统性文献综述厘清这些术语的不同含义。《管理学杂志》。第46卷，第6期，第965–1001页。数字对象标识符：10.1177/0149206320901565

CHEN, C.（2006年）。CiteSpace II：检测和可视化
科学文献中的新兴趋势和瞬态模式。《美国信息科学与技术学会会刊》。第57卷，第3期，第359-377页。数字对象标识符：10.1002/asi.20317

Chen, T.-Y.（2011年）。一种基于多重度量的多准则决策中的乐观与悲观
分析方法，基于Atanassov模糊集。《专家系统与应用》。第38卷，第10期，第12569-12584页。数字对象标识符：10.1016/j.eswa.2011.04.045。

CHEN, T.-Y.（2014a年）。区间值Atanassov直觉模糊环境下的多准则决策分析基于包含关系的LINMAP方法。《国际信息技术与决策期刊》。第13卷，第6期，第1325-1360页。数字对象标识符：10.1142/S0219622014500722。

CHEN, T.-Y.（2014b年）。采用基于似然度的
比较方法的区间值直觉模糊QUALIFLEX方法，用于多准则决策分析。《信息科学》。第261卷，第149-169页。数字对象标识符：10.1016/j.ins.2013.08.054。

CHEN, T.-Y.（2016年）。一种基于包含比较法的多准则决策分析
关于区间值直觉模糊集。《经济的技术与经济发展》。第22卷，第3期，第357-392页。数字对象标识符：10.3846/20294913.2014.989930。

CHEN, T.-Y.（2018年）。一种用于毕达哥拉斯模糊信息多准则决策分析的有效基于相关性的折衷方法。《智能与模糊系统期刊》。 第35卷，第3期，第3529-3541页。数字对象标识符：10.3233/JIFS-18021。CHEN, T.-Y.（2020年）。基于扩展ELECTRE方法进行多准则决策分析时适用于毕达哥拉斯模糊集的新型切比雪夫距离度量。《专家系统与应用》。第147卷，C辑。数字对象标识符：10.1016/j.eswa.2019.113164。COBO, M.J; LÓPEZ-HERRERA, A.G.; HERRERA-VIEDMA, E.; HERRERA, F.（2012年）。SciMAT：一种新的科学制图分析软件工具。《美国信息科学与技术学会会刊》。第63卷，第8期，第1609-1630页。数字对象标识符：10.1002/asi.22688 DEREK, M.（2019年）。利用BERT对讲座进行抽取式文本摘要。arXiv。数字对象标识符：10.48550/arXiv.1906.04165

DESSI, D.; HELAOUI, R.; KUMAR, V.; RECUPERO, D.R.; RIBONI, D.（2021年）。词频-逆文档频率与词嵌入在临床笔记疾病识别中的比较：初步研究。arXiv。数字对象标识符：10.48550/arXiv.2105.09632

DEVLIN, J.; CHANG, M.W.; LEE, K.; TOUTANOVA, K.（2019年）。BERT：用于语言理解的深度双向变换器预训练。载于《2019年计算语言学协会北美分会年会：人类语言技术》论文集。第1卷（长篇和短篇论文），第4171-4186页。明尼阿波利斯，明尼苏达州：计算语言学协会。数字对象标识符：10.18653/v1/N19-1423。

DING Y.（2011年）。科学合作与认可：合著和引文网络的网络分析。《信息计量学杂志》。第5卷，第1期，第187-203页。数字对象标识符：10.1016/j.joi.2010.10.008。FAIRTHORNE, R.A.（1969年）。

用于文献计量描述和预测的经验双曲分布（布拉德福德-齐夫-曼德布罗特）。《文献学杂志》，第25卷第4期，第319–343页。数字对象标识符：10.1108/eb026481

FALAGAS, M.E.; PITSOUNI, E.I.; MALIETZIS, G.A.; PAPPAS, G.（2008年）。PubMed、Scopus、Web of Science 和 Google Scholar 的比较：优势与不足。《美国实验生物学学会联合会期刊》。第22卷第2期，第338–342页。数字对象标识符：10.1096/fj.07-9492LSF。

FLORIANO, C.M.、PEREIRA, V. 和 RODRIGUES, B.E.S.（2022年）。3MO-AHP：通过单目标、多目标或多重目标质量度量降低不一致性的方法。《数据技术与应用》，第56卷，第5期，第645–670页。数字对象标识符：10.1108/DTA-11-2021-0315

GRAUWIN, S.; JENSEN, P. (2011年). 科学机构映射. 科学计量学. 第89卷, 第943–954页.
数字对象标识符: 10.1007/s11192-011-0482-y

GROOTENDORST, M. (2022年). BERTopic：基于类别的TF-IDF方法的神经主题建模. arXiv. 数字对象标识符: 10.48550/arXiv.2203.05794

HARZING, A.W. (2010). 《发表或灭亡》：有效且负责任的引文分析指南. 澳大利亚国家图书馆出版物编目记录. 第1版. 国际标准书号: 978-0-9808485- 0-2

HELDENS, S.; SCLOCCO, A.; DREUNING, H.; WERKHOVEN, B.; HIJMA, P.; MAASSEN, J.; NIEUWPOORT, R.V. (2022年) litstudy：用于文献综述的 Python 软件包. SoftwareX. 第20卷, 第101207期. 数字对象标识符: 10.1016/j.softx.2022.101207

HUG, S.（2003）。比较研究中的选择偏差：不完整数据集案例。《政治分析》第11卷，第3期，第255–274页。http://www.jstor.org/stable/25791732

JEON, H.; KO, H.-K.; LEE, S.; JO, J.; SEO, J.（2022年）。采用两阶段优化的统一流形近似。载于《2022 IEEE可视化与视觉分析》（VIS）。会议论文集。第80–84页。数字对象标识符：10.1109/VIS54862.2022.00025

KOKOL, P.; VOŠNER, H.B.（2018年）。Scopus、Web of Science 和 PubMed 在医学期刊文章资助信息方面的覆盖范围差异。《医学图书馆协会杂志》。第106卷，第1期。数字对象标识符：10.5195/jmla.2018.181. LOTKA, A.J.（1926年）。科学产出的频率分布。《华盛顿科学院学报》。第16卷，第12期，第317–324页。https://www.jstor.org/stable/24529203MAZUREK, J.; PERZINA, R.; STRZAŁKA, D; KOWAL, B.（2020年）。一种用于降低成对比较不一致性的全新逐步（SBS）算法。IEEE Access。第8卷，第135821页–135828页。数字对象标识符：10.1109/ACCESS.2020.3011551。MAZUREK, J.; PERZINA, R.; STRZAŁKA, D; KOWAL, B.; KURAS, P.（2021年）。用于降低成对比较不一致性的迭代算法数值比较。IEEE Access。第9卷，第62553页–62561页。数字对象标识符：10.1109/ACCESS.2021.3074274 McINNES, L.; HEALY, J.; MELVILLE, J.（2018年）。UMAP：用于降维的统一流形近似与投影。arXiv。数字对象标识符：10.48550/arXiv.1802.03426 McLEVEY, J.; McILROY-YOUNG, R.（2015年）。介绍 Metaknowledge：用于信息科学、网络分析与科学学计算研究的软件。《信息计量学杂志》第11卷，第1期，第176–197页。数字对象标识符：10.1016/j.joi.2016.12.005 McQUEEN, J.（1967年）。多变量观测的分类与分析方法。第五届伯克利数学统计与概率研讨会论文集。第1卷，第281–297页。加州大学出版社，伯克利，加利福尼亚州。数学评论编号：MR0214227。https://projecteuclid.org/euclid.bsmsp/1200512992MIRIAM, J.E; URLINGS, B.D.; GERARD, M.H.S.; LEX, M.B.; MAURICE, P.Z.（2021年）。引文偏差及生物医学研究中引文的其他决定因素：来自六个引文网络的发现。《临床流行病学杂志》。第132卷，第71–78页。数字对象标识符：10.1016/j.jclinepi.2020.11.019 OPENAI。（2023年）。GPT-4 技术报告。arXiv。数字对象标识符：https://doi.org/10.48550/arXiv.2303.08774 OTLET, P.（1934年）。《文献论》。布鲁塞尔，Mundaneum，世界宫。https://archive.org/details/OtletTraitDocumentationUgentPEIRSON, B.R.E.; BAKER, A.; SUBRAMANIAN, R.; SINGH, A.; YALUGOTI, Y.（2016年）。Tethne v0.8。可见于http://diging.github.io/tethne/PEREIRA V., COSTA, H.G.（2015年）。应用非线性规划降低层次分析法中的不一致性。《运筹学年鉴》，第229卷，第635页–655页。数字对象标识符：10.1007/s10479-014- 1750-z. PERSSON, O.; DANELL, R.; SCHNEIDER, J.（2009年）。如何使用 Bibexcel 进行各种类型的文献计量分析。《庆祝学术传播研究：献给 Olle Persson 六十岁生日的纪念文集》。5. 9–24。PRITCHARD, A.（1969年）。统计书目学还是文献计量学？《文献学杂志》，第25卷，第4期，第348–349页。RAMKUMAR, A.S.; POORNA, B.（2016年）。运用降维技术进行文本聚类。《国际应用工程研究杂志》。第11卷，第7期，第4770–4774页。数字对象标识符：10.37622/IJAER/11.7.2016.4770-4774 REIMERS, N.; GUREVYCH, I.（2019年）。Sentence-BERT：使用孪生 BERT 网络的句子嵌入。2019年自然语言处理实证方法会议暨第9届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集，第3982–3992页，中国香港。计算语言学协会。< https://aclanthology.org/D19-1410.pdf >RUIZ-ROSERO, J.; RAMIREZ-GONZALEZ, G.; VIVEROS-DELGADO, J.（2019年）。软件综述：Scientopy，一种用于分析科学出版物主题趋势的科学计量工具。《科学计量学》。第121卷，第1165–1188页（2019年）。数字对象标识符：10.1007/s11192-019-03213-w

SENGUPTA, I.N.（1992）。文献计量学、信息计量学、科学计量学与图书馆计量学：概述。第42卷，第2期，第75-98页。数字对象标识符：10.1515/libr.1992.42.2.75 SRISAWAD, S.; LERTSITTIPHAN, K.（2021）。泰国多学科期刊文章中影响引文的参考文献选择之文献计量分析。科学计量学研究杂志。第10卷，第3期，第288-296页。数字对象标识符：10.5530/jscires.10.3.45 THOR, A.; MARX, W.; LEYDESDORFF, L.; BORNMANN, L.（2016）。介绍被引参考文献浏览器（CRExplorer）：一种用于参考文献出版年份光谱分析并进行被引参考文献标准化的程序。信息计量学杂志。第10卷，第2期，第503-515页。数字对象标识符：10.1016/j.joi.2016.02.005 VAN ECK N.J.; WALTMAN L.（2010）。软件综述：VOSviewer，一款用于文献计量映射的计算机程序。科学计量学。第84卷，第2期，第523-38页。数字对象标识符：0.1007/s11192-009-0146-3 VAN ECK, N.J.; WALTMAN, L.（2014）。CitNetExplorer：一种用于分析和可视化引文网络的新软件工具。信息计量学杂志。第8卷，第4期，第802-823页。数字对象标识符：10.1016/j.joi.2014.07.006 YADAV, D.; LALIT, N.; KAUSHIK, R.; SINGH, Y.; MOHIT; DINESH; YADAV, A.K.; BHADANE, K.V.; KUMAR, A.; KHAN, B.（2022）。文本摘要技术及其在健康领域应用的定性分析。计算智能与神经科学。第2022卷，文章编号3411881，共14页。DPI：10.1155/2022/3411881 ZHANG, W.; JU, Y.; LIU, X.（2017）。基于Shapley模糊测度和区间值犹豫模糊语言数的多准则决策分析。计算机与工业工程。第105卷，第28-38页。数字对象标识符：10.1016/j.cie.2016.12.046。ZHANG, J.; ZHAO, Y.; SALEH, M.; LIU, P.J.（2019）。PEGASUS：通过提取的间隙句子进行预训练，以实现抽象式摘要。ICML'20：第37届国际机器学习会议论文集机器学习，2020年7月，文章编号：1051，第11328-11339页。<https://dl.acm.org/doi/pdf/10.5555/3524938.3525989 >