·医学信息研究·

基于双聚类算法探测引文分析的知识基础及研究前沿

侯跃芳 王 潇

（中国医科大学医学信息学院 沈阳 110122） （北京京都儿童医院病案室 北京 102208）

〔摘要〕 基于双聚类算法对引文分析相关的高被引文献及其来源文献同时聚类，从高被引论文（知识基础）出发，依据聚类模块选择与高被引论文簇关系最为密切的来源文献代表研究前沿，总结引文分析的知识基础和研究前沿，从而了解专题发展脉络，为科学研究提供参考方向。

〔关键词〕 双聚类；引文分析；知识基础；研究前沿

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2016.01.012

Exploration of Knowledge Bases and Research Fronts of Citation Analysis Based on Bi－clustering Algorithm *HOU Yue－fang, School of Medical Informatics, China Medical University, Shenyang 110122, China; WAND Xiao, Medical Records Department, Beijing Jingdu Children's Hospital, Beijing 102208, China*

〔**Abstract**〕 The paper clusters highly cited documents related to citation analysis and their source documents based on bi－clustering algorithm. Starting from highly cited documents (knowledge bases), it selects source documents most closely related to the highly cited document cluster based on the clustering module to represent the research fronts, summarizes the knowledge bases and research fronts of citation analysis, understands the development of special topics and provides a reference for scientific research.

〔**Keywords**〕 Bi－clustering; Citation analysis; Knowledge base; Research front

1 引言

引文分析自问世以来，在信息计量学界得到了广泛应用，其相关研究也日趋完善和多样化。探测引文分析的知识基础及研究前沿，有利于了解专题发展脉络，把握科技创新前沿，为科学研究提供参考方向。

1965 年普赖斯基于引文分析最早提出“研究前沿”的概念，指出科学家积极引用的文献代表了该领域的前沿[1]。1994 年 Persson 则明确区分研究前沿和知识基础两个概念，其中知识基础由同被引文献簇表示，研究前沿则是由文献耦合方法生成的、与知识基础有引证关系的文献群，指出引证文献更能代表当前的研究[2]。无论是同被引分析还是耦合分析，都是引用传统聚类算法在单一维度上的聚类，不能同时对被引文献和引用文献聚类；而且相对于高被引论文，引用文献簇众多，如何选择有代表性的来源文献成为研究瓶颈。双聚类算法的提出为解决这一问题提供了可能。

Hartigan 于 1971 年提出双聚类的概念，即同时聚类（Simultaneous Clustering）[3]，实现了在对象及

〔修回日期〕 2015－11－11

〔作者简介〕 侯跃芳，博士，副教授，发表论文 30 余篇。

其属性两个方向上的同时聚类，通过提取对象及其属性的联合信息发现潜在的局部模式。目前，双聚类分析方法在生物信息学中得到广泛应用，但将双聚类用于文献研究的比较少，如期刊 - 关键词的双聚类[4]、主题词 - 论文的双聚类[5]。本校崔雷教授研究组率先利用双聚类算法探测学科前沿及知识基础（以 h 指数专题为例），验证了此方法的先进性和可行性[6]。本论文基于双聚类算法对引文分析相关的高被引文献及其来源文献同时聚类，从高被引论文（知识基础）出发，依据双聚类图谱中的聚类模块，选择与高被引论文簇关系最为密切的来源文献代表研究前沿，得到引文分析相关研究的知识基础和研究前沿。

2 研究方法

2.1 检索引文分析相关文献

利用 SCI 数据库（1900 年至今）和 SSCI 数据库（1983 年至今）检索引文分析的相关研究，检索式为：TS = “citation analysis” OR “citation analyses” OR “co - citation *” OR “cocitation *” OR “co - cited” OR “bibliographic coupling” OR “citation link *” OR “citation network *” OR “citation ranking *”。选择“article”文献类型，获得相关文献 2 573 篇。检索时间为 2015 年 1 月。

2.2 引文频次统计及矩阵生成

利用 BICOMB 软件[7]统计引文频次。根据 h 指数的原理，截取 55 篇高被引文献（h = 54），见表 1，构建高被引论文 - 来源文献矩阵。

表 1 引文分析相关的 55 篇高被引论文

序号	高被引论文	被引频次	序号	高被引论文	被引频次
1	SMALL H. J AM SOC INFORM SCI, 1973, 24: 265	301	29	Borner K. ANNU REV INFORM SCI, 2003, 37: 179	67
2	GARFIELD E. SCIENCE, 1972, 178: 471	249	30	MERTON RK. SCIENCE, 1968, 159: 56	67
3	Hirsch JE. P NATL ACAD SCI USA, 2005, 102: 16569	223	31	SMALL H. SCIENTOMETRICS, 1985, 7: 391	67
4	WHITE HD. J AM SOC INFORM SCI, 1981, 32: 163	196	32	Boyack KW. SCIENTOMETRICS, 2005, 64: 351	65
5	White HD. J AM SOC INFORM SCI, 1998, 49: 327	183	33	CULNAN MJ. MANAGE SCI, 1986, 32: 156	65
6	PRICE DJD. SCIENCE, 1965, 149: 510	174	34	Brin S. COMPUT NETWORKS ISDN, 1998, 30: 107	64
7	MCCAIN KW. J AM SOC INFORM SCI, 1990, 41: 433	148	35	PINSKI G. INFORM PROCESS MANAG, 1976, 12: 297	64
8	GARFIELD E. SCIENCE, 1955, 122: 108	134	36	SEGLEN PO. J AM SOC INFORM SCI, 1992, 43: 628	62
9	KESSLER MM. AM DOC, 1963, 14: 10	134	37	BRAAM RR. J AM SOC INFORM SCI, 1991, 42: 252	61
10	SMALL H. SCI STUD, 1974, 4: 17	134	38	MORAVCSIK MJ. SOC STUD SCI, 1975, 5: 86	61
11	MACROBERTS MH. J AM SOC INFORM SCI, 1989, 40: 342	132	39	KAMADA T. INFORM PROCESS LETT, 1989, 31: 7	60
12	Garfield E. CITATION INDEXING, 1979	113	40	Kleinberg JM. J ACM, 1999, 46: 604	60
13	Moed H. CITATION ANAL RES EV, 2005	100	41	SMALL H. SCIENTOMETRICS, 1985, 8: 321	60
14	SMALL HG. SOC STUD SCI, 1978, 8: 327	97	42	White HD. J AM SOC INF SCI TEC, 2003, 54: 423	60
15	BRAAM RR. J AM SOC INFORM SCI, 1991, 42: P233	92	43	KUHN T. S. STRUCTURE SCI REVOLU, 1962	59
16	Garfield E. CITATION INDEXING IT, 1979	91	44	Kuhn T. STRUCTURE SCI REVOLU, 1970	59
17	Price DJS. LITTLE SCI BIG SCI, 1963	91	45	Meho LI. J AM SOC INF SCI TEC, 2007, 58: 2105	59

续表 1

18	Seglen PO. BRIT MED J, 1997, 314: 498	88	46	Cronin B. CITATION PROCESS ROL, 1984	57
19	Wasserman S. SOCIAL NETWORK ANAL, 1994	85	47	CULNAN MJ. MIS QUART, 1987, 11: 341	56
20	SMITH LC. LIBR TRENDS, 1981, 30: 83	82	48	FREEMAN LC. SOC NETWORKS, 1979, 1: 215	56
21	Small H. J AM SOC INFORM SCI, 1999, 50: 799	80	49	MCCAIN KW. J AM SOC INFORM SCI, 1991, 42: 290	56
22	Barabasi AL. SCIENCE, 1999, 286: 509	78	50	PERSSON O. J AM SOC INFORM SCI, 1994, 45: 31	56
23	Ahlgren P. J AM SOC INF SCI TEC, 2003, 54: 550	71	51	Albert R. REV MOD PHYS, 2002, 74: 47	55
24	Garfield E. JAMA – J AM MED ASSOC, 2006, 295: 90	71	52	Crane D. INVISIBLE COLL DIFFU, 1972	55
25	MacRoberts MH. SCIENTOMETRICS, 1996, 36: 435	71	53	Redner S. EUR PHYS J B, 1998, 4: 131	55
26	WHITE HD. ANNU REV INFORM SCI, 1989, 24: 119	71	54	Chen CM. J AM SOC INF SCI TEC, 2006, 57: 359	54
27	GARFIELD E. SCIENTOMETRICS, 1979, 1: 359	70	55	HUMMON NP. SOC NETWORKS, 1989, 11: 39	54
28	Egghe L. SCIENTOMETRICS, 2006, 69: 131	69			

2.3 双聚类分析

将高被引 – 来源文献矩阵导入 gCLUTO 软件进行双聚类分析。聚类方法选择重复二分法（Repeated Bisection），最优化函数选择 I2，相似系数选择余弦函数（Cosine）。以矩阵行的聚类代表高被引论文的聚类，列的聚类代表来源文献聚类。gCLUTO 中计算聚类后各类的相似性指标，作为衡量聚类质量的参考。ISim 表示类内对象间的平均相似度，ISdev 为 ISim 的标准差，ESim 表示类间平均相似度，ESdev 为 ESim 的标准差。类内相似度越高，类间相似度越低，说明聚类效果越好。除了衡量聚类内部质量的相似性指标，gCLUTO 还考虑了聚类的外部特征：描述属性和区分属性。某一描述属性的百分比数据为该属性可以解释的类内相似度百分比，某一区分属性的数据为该属性可以解释的该类与其他对象间相异度的百分比。一般情况下，描述属性和区分属性之间有大量重叠，唯一的区别是区分属性的百分比数据一般要比描述属性小。原因是某个类别中的一些描述属性也会存在于不属于该类别的少量对象中。这些描述性特征可以为总结聚类的特征及研究前沿提供有力的线索。高被引文献代表引文分析的知识基础，描述性特征（来源文献）代表研究前沿。gCLUTO 还将聚类结果用可视化图谱及山峰图显示。山峰的体积与类群包含的对象数量成比例，高度与类内相似性成比例。类内相似性越大，则山峰越陡峭。这种多维标度可视化方法展示了双聚类的效果，可以据此推测将对象聚为几类更理想。

3 结果与分析

综合 gCLUTO 得出的相似性指标及可视化山峰图判断，见图 1，高被引论文聚为 8 类的效果比较理想。可视化矩阵，见图 2。显示了高被引文献和来源文献两个维度的聚类结果。行聚类表示高被引文献聚为 8 类，列聚类表示对来源文献的聚类，颜色深浅代表文献被引频次的高低。结合双聚类图谱中的聚类模块分布及各类描述属性、区分属性特征（表略），对各类聚集模块中高被引论文以及与之关系密切的来源文献的内容进行总结，得出引文分析各个研究方向的知识基础与研究前沿。

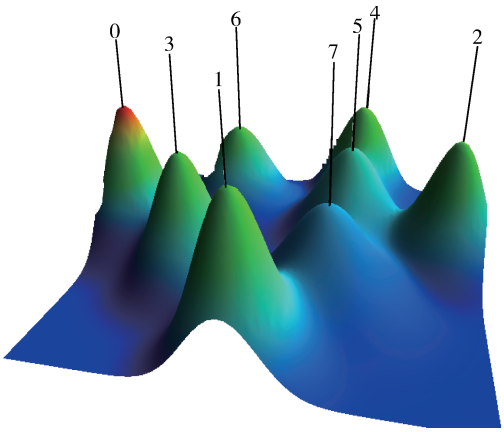

图 1 双聚类可视化山峰

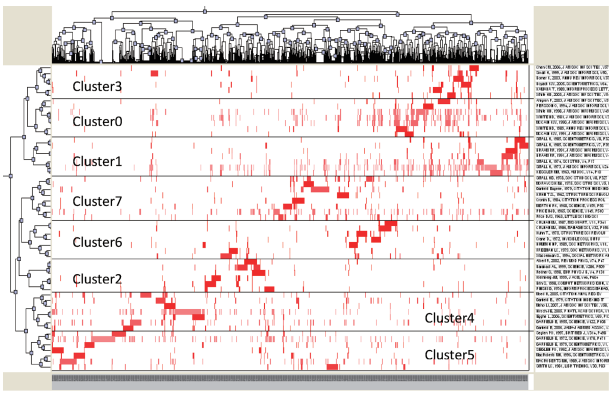

图 2 双聚类可视化矩阵

3.1 类 0—共引分析溯源、应用与概述

属此类的有表 1 中高被引论文 4，5，7，23，26，49，50。知识基础：作者共被引分析方法。1981 年 White HD 提出科学结构的文献测量方法——作者共被引分析。1990 年 Mccain KW 将该方法归纳为 6 步，即选择作者、检索同被引频次、构造同被引矩阵、转化为皮尔逊相关系数矩阵、多元分析、解释结果及效度分析。该模式以 SPSS、SAS 等统计学软件为工具，利用聚类分析、多维定标和因子分析等技术，以映射地图的方式定量地刻画科学结构，寻找科学范式。2003 年 Ahlgren P 提出皮尔逊相关系数不是测度相似性的最佳选择。研究前沿：共引分析的深入应用及技术探讨。2010 年 Uysal OO 采用文献计量学分析来研究商业伦理。Georgi C 将其应用于物流和供应管理。2008 年 Sugimoto CR 运用字段共引分析图书馆信息科学和管理信息系统的关系。van Eck NJ 比较两种知识图谱的绘制技术（多维定标分析和 VOS）。

3.2 类 1—共引关系、引文图谱构建与应用

属此类的有表 1 中高被引论文 1，9，10，15，31，37，41。知识基础：文献耦合、共引分析及共词分析方法。1963 年 Kesssler MM 提出了文献耦合。1973 年 Small H 提出了文献共引的概念和共引分析的方法，定义共引强度来测量论文之间的共引程度，认为共引是测量两篇文献相关度的新方法。1974 年 Small H 提出了著名的圆环模型，形象地表示了文献间存在的双引、三引等共引关系，Small H 是共引分析的开创者。研究前沿：引文图谱构建方法及利用引文方法分析学科结构。Moya – Anegon F 及 Rafols I 构建大型共引图谱来定位科学研究，进行技术改进。1993 年 MILMAN BL 应用引文和共引分析研究化工领域的信息和知识流程。

3.3 类 2—引文网络

属此类的有表 1 中高被引论文 22，34，35，40，51，53。知识基础：引文网络。2002 年 Albert R 总结了复杂网络领域的最新进展。1999 年 Barabasi ALB 研究了复杂网络的各种拓扑结构、网络功能及演化规律等，重点是网络的动力学，包括网络中的博弈演化及统计。讨论了主要模型和分析工具，包括随机图、小世界和无标度网络。Kleinberg JM 介绍了超链环境中的权威网页和枢纽页。研究前沿：引文网络的结构探讨及深入应用。2009 年 Radicchi F 利用引文网络探讨科学研究者的知识扩散和科学家的排名。2012 年 Medina CMC 应用引文网络来鉴别核心期刊。Leicht EA 探讨随时间演变的引文网络大型结构。

3.4 类 3—引文分析可视化图谱

属此类的有表 1 中高被引论文 21，29，32，39，42，54。知识基础：引文分析的可视化图谱。1989 年 KAMADA T 提出绘制一般无向图的一种算法。1999 年 Small H 探讨科学引文可视化图谱。2003 年 Borner K 等介绍了知识可视化的流程，用几

种新兴的可视化工具和方法描绘了科学计量学的主流领域，得出引文分析是科学计量学的主要领域。研究前沿：引文分析可视化图谱的应用和方法改进。2010 年 Takeda Y 和 Vargas－Quesada B 利用引文网络图谱展现科学领域的基本结构和演变，有具体方法改进。还有学者改进引文分析图谱的运算方法及开发绘制图谱的软件。

3.5 类 4—引文分析指标与应用

属此类的有表 1 中高被引论文 3，8，13，16，24，28，45。知识基础：引文分析指标及评价。2006 年 Garfield E 阐述了期刊影响因子的历史和意义。2005 年 Hirsch JE 提出 h 指数，可用于评估研究人员的学术研究水平。2006 年 Egghe L 提出了 g 指数，是对 h 指数的完善。1955 年 Garfield E 创立了 SCI，阐明文献之间的引用关系本质上是知识的联系，也就是知识的传递流动过程，为研究人员成果的评估奠定了基础。研究前沿：引文分析指标的深入应用和评价。2009 年以来，研究人员评价引文指标的使用和滥用情况，利用引文指标评估科研绩效，度量期刊网点运营管理情况等。

3.6 类 5—引文分析评价的应用与问题

属此类的有表 1 中高被引论文 2，11，18，20，25，27，36。知识基础：引文分析评价的应用及存在问题。1972 年 Garfield E 提出利用引文分析评价期刊，通过引文的频次和影响来排序期刊。1979 年 Garfield E 阐述了引文索引理论和方法，以及引文分析在自然科学和人文社会科学领域的应用。通过引文分析对期刊以及研究成果等进行评价，提出了科学引文索引是一个能评价研究绩效、促进科学进步的新工具。1989 及 1996 年 MacRoberts MH 提出了引文分析存在的问题。研究前沿：引文分析评价研究绩效的深入探讨及引文分析存在问题的解决。2000 年 Meho LI 研究得出资深教师研究成果的同行评议和引文排名是相反的，提出引文排序可以作为教师成果评价的有效指标。近年也有学者针对引文分析存在的问题，提出替代影响因子的指标或综合性影响指标。

3.7 类 6—引文分析揭示学科发展

属此类的有表 1 中高被引论文 19，33，44，47，48，52，55。知识基础：引文分析展现热门学科发展结构。1986－1987 年 CULNAN MJ 利用共引分析展现管理信息系统的知识发展。1989 年 HUMMON NP 利用引文网络的连通性探讨 DNA 研究的理论发展。研究前沿：利用引文分析发现科技研究前沿。近年有研究探讨新知识的发展模式、算法理论，或寻找不同学科的研究前沿（如商业伦理方面）。

3.8 类 7—引用理论探讨

属于此类的有表中高被引论文 6，12，14，17，30，38，43，46。知识基础：科学界引用过程的理论探讨。1963 年 Price DJS 研究了科学的指数型发展和科学家的作用。1965 年他研究了科学论文之间的引证和被引证关系，以及由此形成的引证网络。1968 年 MERTON RK 论述科学界的马太效应。研究前沿：引用理论的再探讨及其应用。1998 年 Leydesdorff L 从科学交流的文化角度来解构引文分析。2010 年 Lin TY 提出绘制社会行为研究的知识结构图谱。

4 结论

本研究利用双聚类方法探测了引文分析相关研究的知识基础和学科前沿，进一步验证了此种方法的可行性和便利性。后序研究中，可以将双聚类分析推广至各学科发展的文献评价领域中。这对科学研究人员科技创新，甚至对各国政府制定科技发展战略都具有重要意义。双聚类算法在对高被引文献聚类的同时，也对来源文献进行聚类，增加类内高被引文献和来源文献的相关性，即知识基础与研究前沿的相关性。而且根据双聚类图谱中的聚类模块选择具有代表性的来源文献，可解决根据大量来源文献难以总结研究前沿的瓶颈问题。

也有研究者利用 CiteSpace 得到的突发词及其论文的引文探测学科前沿及知识基础[8]。利用突发词

展现学科前沿方便快捷，但是有些突发词比较宽泛，不能准确表达前沿内容。而且突发检测法易受噪音文献影响，利用突发词作为聚类标签也存在争议。本研究依据 h 指数的原理计算引用频次阈值，提取高被引论文，验证其可行性，扩展 h 指数的应用。但是在双聚类分析过程中，仅对高被引文献及其对应的来源文献进行分析，可能会使结果不全面。另外，对于聚类数目和效果的选择取决于感官判断，可能会造成结果偏倚。

参考文献

1 Price DD. Networks of Scientific Papers [J]. Science, 1965, (149): 510 - 515.

2 Persson O. The Intellectual Base and Research Fronts of JASIS 1986 - 1990 [J]. J AM SOC INFORM SCI, 1994, 45 (1): 31 - 38.

3 Hartigan JA. Direct Clustering of a Data Matrix [J]. J Am Stat Assoc, 1972, 67 (337): 123 - 129.

4 于跃，徐志健，王珅，等. 基于双聚类方法的生物医学信息学文本数据挖掘研究 [J]. 图书情报工作，2012，56 (18)：133 - 136.

5 李范，李敏，王丽，等. 利用共词分析挖掘国际护理信息学研究热点 [J]. 医学信息学杂志，2014，35 (9)：48 - 53.

6 方丽，崔雷. 利用双聚类算法探测学科前沿及知识基础——以 h 指数研究领域为例 [J]. 情报理论与实践，2014，37 (11)：55 - 60

7 崔雷，刘伟，闫雷，等. 文献数据中书目信息共现挖掘系统的开发 [J]. 现代图书情报技术，2008，(8)：70 - 75.

8 方丽，赵悦阳，崔雷. 利用突发检测算法探测学科前沿及知识基础 [J]. 医学信息学杂志，2014，35 (10)：49 - 54.

(上接第 45 页)

3 结语

本系统为各类教学资源提供安全快捷的存储管理，为各种使用者提供丰富多样的学习资源，为教学管理者提供全面系统的效果评价分析，从而提高教学质量，促进教学的多元化发展，也为教师和学者提供一个资源共享和学术交流的平台。网络教学资源管理平台建设是目前高校信息化建设的重要环节，面对快速发展的网络教学资源，只有不断地更新和完善系统资源，才能保证平台真正服务于广大教师学者。

参考文献

1 吴美娇，项国雄. 国家精品课程网络教学资源现状分析与优化 [J]. 现代远程教育研究，2009，(2)：39 - 44.

2 张梅，李树民，唐品，等. 基于 BlackBoard 平台的医学信息检索精品课程建设 [J]. 医学信息学杂志，2010，31 (8)：93 - 95.

3 王涛，裘国永，宋伟，等. 基于 CMS 的精品课程网站建设研究与实践 [J]. 现代教育技术，2011，21 (6)：120 - 122.

4 张稚鲲，李文林，郝桂荣. 文献检索网络课程教学模式探索与实践 [J]. 医学信息学杂志，2011，32 (12)：80 - 83.

5 郭广军，谢东，李魏豪. 基于 CMS 的网站系统开发技术研究及应用 [J]. 计算机工程与设计，2010，31 (11)：2500 - 2502.

6 耿璐，聂足. 基于 CMS 的企业网站的设计与实现 [J]. 计算机工程与设计，2009，30 (2)：351 - 357.

7 管华，李禹生，徐军利，等. 基于网络教学资源平台的个性化自主学习研究 [J]. 计算机教育，2012，(6)：94 - 98.

8 张家贵，罗龙涛. 基于云计算理念构建数字化教学资源平台 [J]. 现代教育技术，2011，21 (3)：102 - 104.

9 谭立球，费耀平，李建华，等. 多网站内容管理系统的设计和实现 [J]. 计算机应用，2004，24 (11)：4 - 6.

10 王聪，房国栋. 高校精品课程网络教学资源构建模式比较研究 [J]. 现代教育技术，2010，20 (9)：46 - 49.

11 李海宝，任常愚. 基于网络公共空间的辅助教学平台的构建 [J]. 中国远程教育，2014，(4)：84 - 88.