基于可视化的作者同被引技术的发展

宋丽萍 徐引篪

(中国科学院文献情报中心，北京 100080)

摘要 自1981年作者同被引技术(ACA)开发以来，作为一种揭示科学结构行之有效的方法，被学术界认可并广为采用，但烦琐的程序、计算中存在的问题严重阻碍了其发展。信息可视化浪潮为ACA的发展带来了契机。本文重点阐述了将信息可视化技术用于作者同被引分析的成果。

关键词 作者同被引 网络寻址定位 作者链接

Information Visualization and the Development of ACA

Song Liping and Xu Yinchi

(*Library of Chinese Academy of Sciences, Beijing* 100080)

Abstract As an effective method for displaying intellectual structure, Author Co-citation Analysis(ACA) was widely used and generally accepted, but it was, and is still labor intensive and time consuming work. Information Visualization(IV) provides a new tool for ACA. In this article, it introduces the results applying IV into ACA.

Keywords ACA, PFNETs, authorlink.

1981年，美国费城德瑞克赛大学(Drexel University)成为作者同被引技术(Author Co-citation Analysis，简称ACA)诞生的摇篮。H.D.White博士和K.W.Mccain博士以同被引频次越高，则作者学术相关性越强作为分析的大前提，通过世界39位情报科学作者的同被引分析描述了情报科学结构[1]，因而引起情报界特别是文献计量领域的瞩目。该文以SPSS(美国社会科学统计软件包)为工具，采用聚类分析(Cluster Analysis)、多维定标(Multidimensional Scaling)和因子分析(Factor Analysis)定量刻画了情报科学学术框架，为ACA分析提供了良好的范例。1990年，Mccain将ACA的程序归整为选择作者、检索同被引频次、构成同被引矩阵、转化为皮尔逊相关系数矩阵、多元分析和解释结果等几个步骤，人们称其为传统ACA或冠之以德瑞克赛模式。自1981年以来的20年里，人们沿用传统ACA方法描述不同的领域或同一学科的不同发展阶段，甚至在新加坡南洋技术学院近来开发的网络引文检索系统Pubsearch中也采用了该方法。然而2003年，传统ACA遇到了挑战。

1 鲁索的质疑与技术的推动

2003年，堪称ACA发展史上具有划时代意义的一年。一方面是著名比利时情报计量学家鲁索(R. Rousseau)等3人在JASIST(Journal of the American Society for Information Science and Technology)上发表《皮尔逊相关系数与同被引相似性计算》一文，文中对传统ACA采用皮尔逊相关系数度量作者相关性提出质疑，认为在作者相似性计算中，皮尔逊相关系数并非最佳选择，因此沿用长达20年之久的传统ACA受到了挑战。而另一方面，近年来信息可视化

收稿日期：2004年5月18日

作者简介：宋丽萍，女，中科院文献情报中心博士研究生，天津师范大学讲师。徐引篪，女，研究馆员，博士生导师。

技术渗透到情报科学领域，也为 ACA 的发展提供了契机。1997 年美国肯塔基州立大学的 Linxia 即已开始尝试将自组织映射技术（Self-organization Map，简称 SOM）应用到 ACA 分析中；1999 年和 2000 年，英国的 C.Chen（后转入美国德瑞克赛大学）将潜在语义标引（Latent Semantic Indexing）和网络寻址定位（Pathfinder Network Scaling，简称 PFNETs）融入 ACA 中；2000 年，ACA 技术的开发者之一 White 以及 Buzydlowski，Linxia 三人提出采用 PFNETs 产生同被引作者图，并用于文献检索的设想，ACA 在其发源地——德瑞克赛大学又获得了新的生机。由于内在发展张力、外在技术推动，ACA 突破了传统模式的困囿，在描述学科结构方面取得了重大进展，并开辟了新的应用领域——主题检索。

2 ACA 在描述科学结构方面的进展

描述科学结构是 White 开发 ACA 的初衷，ACA 实则是一个学科的微缩景观。继 1981 年之后，White 于 1998 年沿用传统 ACA 对情报科学结构进行了描述，并于 2003 年采用 PFNETs 对 1998 年的同一数据进行了第二次分析。

2.1 传统 ACA 分析结果

1998 年，White 及 Mccain 对 1972～1979 年，1980～1987 年，1988～1995 年 24 年间 SSCI 进行统计，样本范围为《情报科学技术年评》、《科学计量学》、《美国情报会志》、《电子图书馆》等 12 种期刊，统计对象为 12 种期刊中高频被引的作者，统计筛选出 120 位情报科学作者，然后采用传统 ACA 技术对上述作者进行同被引聚类分析，结果如下[2]：

- 二维体系图犹似澳大利亚版图：沿海地区迅速发展，中部地区人口稀少。
- 聚类结果将情报科学大致分为两大阵营：文献计量（包括引文分析），情报检索。两大阵营泾渭分明。
- 就数量而言，检索学家占据优势。
- 分析中并未出现能够对学科发展起导向作用的核心作者，或者核心作者集团。

2.2 PFNETs 分析结果

就鲁索提出的质疑，White 做出的反应一是在 JASIST 上发表了《作者同被引与皮尔逊相关系数 r》一文[3]，回答了鲁索的问题，并就采用皮尔逊相关系数的渊源做出解释；二是随即发表《网络寻址定标与情报科学范式的再划分》[4]，文中采用网络寻址定标对 1998 年同一数据进行了分析。

文中，White 首先采用 KNOT 软件将 121 位作者所产生的 $121 \times 120 \div 2 = 7\,260$ 个组合缩减到较为重要的 126 个。由于 KNOT 的可视性较差，所以采用用于大型网络分析、可视性较强的 Pajek 软件。但是 Pajek 用相同的点代表作者，而 White 则采用大小不同的点代表不同的作者，以突出作者在学科中的不同影响力及其重要程度。结果如图 1 所示。

由图 1 可以得出结论：

- 4 位情报科学的领袖人物脱颖而出：G.Salton，E.Garfield，F.W.Lancaster，其次是 D.Price。
- 以领袖人物为核心聚集成情报科学作者链：Markey-Bates-Berlkin-Saracevic-Salton-Lancaster-Garfield-Price-Brookes。并围绕这些作者构成了 1972～1995 年间情报科学研究范式：从 Markey 到 Saracevic 的作者，致力于非实验文献检索系统研究；Salton 和 Garfield 各自统领两个最大的学术集团，即检索学家和文献计量（含引文分析）学家；Lancaster 成为一般理论集团的焦点人物。
- Lancaster 作为最核心的情报学家，成为连接检索集团与文献计量集团的纽带。

2003 年的分析结果中反映了多个学术集团并存的格局，并突出了核心人物，这样更加符合情报科学的认知结构。显然，2003 年 PENETs 的分析结果优于 1998 年的传统 ACA 的分析结果。

2.3 PFNETs 的优势

客观而言，传统 ACA 需要大量的计算与绘图操作。分析者首先必须通过各种来源确定能够覆盖一个学科各个分支的作者集合，进而通过分析程序，并依赖于支持因子分析、多维定标以及聚类分析的统计工具（例如 SPSS），通过多维图观察相似性而形成集团，同时借助统计方法确定作者的重要性。这种工作流程不仅烦琐和复杂，并且转换过程很可能造成数据失真而影响聚类结果。结果不理想时，研究者甚至需要按照作者的位置人工聚类。一言以蔽之，传统 ACA 很大程度上需要人为干预。

在 ACA 问世的 20 年内，虽则作为一种揭示科学结构的方法得到学术界的认可，但烦琐的数据搜集，计算中存在的矩阵对角线值设定问题、皮尔逊相关系数转换问题严重阻碍了其广泛应用，于是人们开始探讨用其他相似性计算方式替代传统的皮尔逊

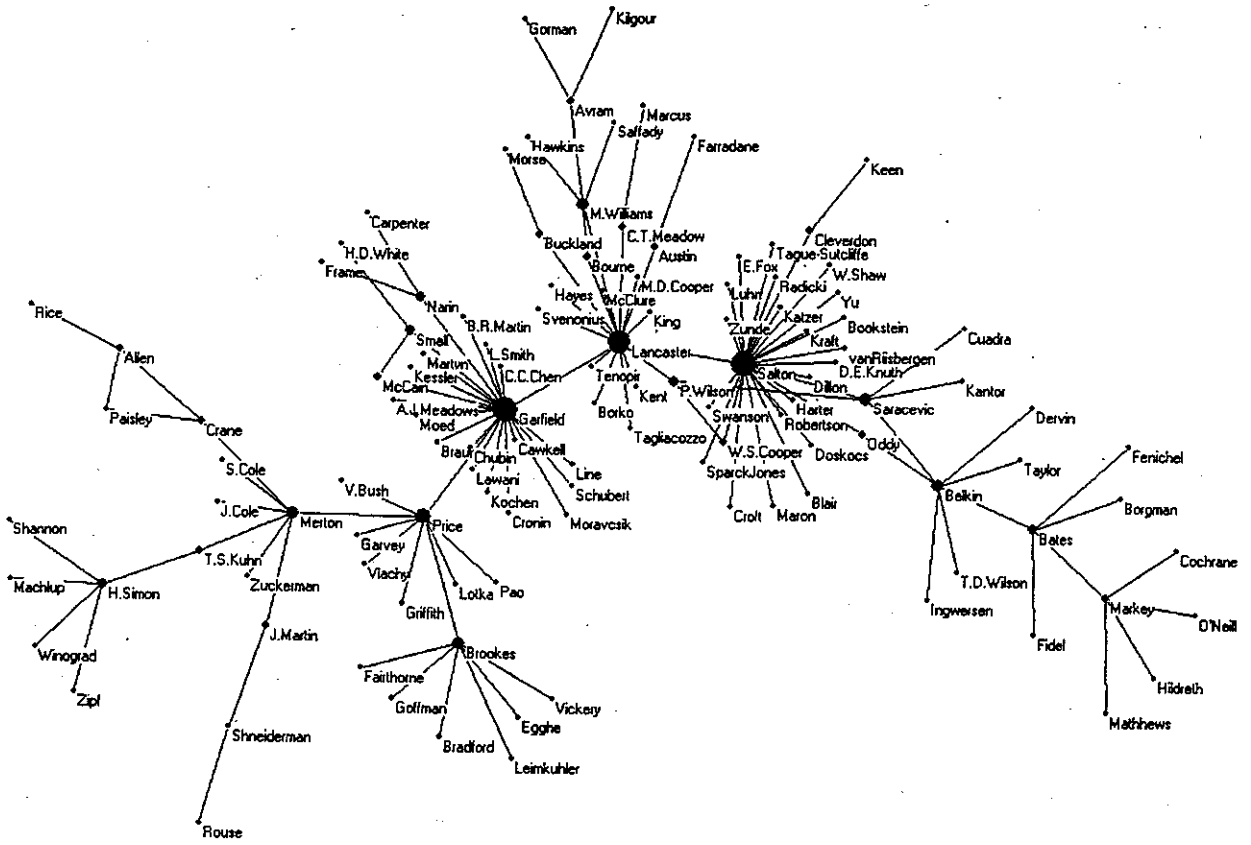

图 1　121 位作者原始同被引频次网络寻址定标(引自文献[1])

相关系数。PFNETs 就是人们作出的尝试之一。PFNETs 起源于 1990 年美国心理学家 Schvaneveldt 的认知心理学语义关系研究,它与社会网络分析具有共同的数学模型[5],将 PFNETs 引入 ACA 的同时,也就将认知心理语义分析、社会网络分析统计和可视化软件的优势借鉴到 ACA 中来,因此为传统 ACA 注入了活力。PFNETs 方法采用原始同被引频次,将作者视为节点,而假设节点间由加权的路径相连,权值为作者的同被引频次,并且仅显示节点间最短路径。在图中,与许多作者相连的宛若"非正式交流网络"中的超星,他们控制着学科研究的走向。其余作者以超星为核心形成不同的研究范式,进而构成学科结构全景。相反,如果一个领域缺乏领袖级的人物,PFNETs 则呈现出相对松散的状态。

PFNETs 计算快捷并具有很强的可视性,正如 Pajek 设计原则中所强调的:本软件宗旨之一就是为用户提供一个强大的可视化工具。其次,许多作者与某个核心作者同被引强度都很高时,专业分支自动形成,而无须单独的聚类程序。因而比起采用皮尔逊系数,利用原始同被引频次的 PFNETs 结果更为丰富。尤为重要的是 PFNETs 减少了传统 ACA 模式的复杂性,并且结果更为可信,因而在现今 ACA 分析中倍受推崇。

3　ACA 在主题检索方面的新进展

1976 年,Clevendon 指出:作者一定程度上是其研究主题的代名词。早在 20 世纪 90 年代初期,Mccain 就曾设想将 ACA 用于综述某一学科的发展。2000 年,White 提出作为聚类的副产品,与已知作者从事同一主题研究的相关作者一览无遗,因此可以将 ACA 应用拓展到主题检索,从而体现了将同被引用于主题检索思想的萌芽。但是由于技术等方面的原因,时至今日才梦想成真。

德瑞克赛大学的 H.D.White 是 ACA 技术的开发者,他带领由 Xia Lin,Jan Buzydlowski 组成的研究小组正在开展实时环境下 ACA 绘图及主题检索的研究。这种基于 Web 的实验系统名为 AuthorLink。这一系统是基于信息可视化技术实现的。信息可视化(Information Visualization)这一概念是由计算机图形协会成员 Mccormick 在 1987 年提出的,其宗旨是在计算机协助下,通过对数据可见的、交互的表示,

从而洞察数据，发现信息[6]。目前可视化已经应用到信息检索领域。具体而言，就是把文献信息、用户提问、各类情报检索模型以及利用检索模型进行信息检索的过程中，不可见的内部语义关系转换成图形，将高维性的数据库，在一个二维或三维的可视化空间中显示出来，使得不可见的关系用可见的方式表达出来。可视化空间中的普通对象表现为空间中的点，对象间的关系则表现为点间连线。可视检索的关键是降低高维向量空间的维数。减少维数涉及到的算法大概包括：主成分分析法、多维定标技术、自组织映射技术、分层聚类法和网络寻址定标法。可视化技术用于 ACA，将作者表示为二维空间的点，而用连线表示作者间关系，在此基础上划分为不同学术分支，并转化成可视化情报检索界面(VIRIs)。目前，AuthorLink 在德瑞克塞大学的网站上运行。

3.1 AuthorLink 的功能

当输入一个作者名时，用户从该系统得到的将不仅是一个作者的信息，而是与该作者高频同被引的 24 位作者，以及基于同被引强度以图的形式展示的作者间相互关系[7]。目前 AuthorLink 能够通过美国情报科学研究所提供的 1988 ~ 1997 年间 AHCI (艺术与人文引文索引)数据库实时生成交互链接图，共 126 万条记录。其应用方法如下：

- 输入一位作者名，系统将显示与其同被引的前 24 位作者名；
- 点击 Map It Now，系统将在几秒内自动实时生成同被引作者图(Instant Author Co-citation Map)；同时点击 Show The Numbers，则立即在作者连接线上显示同被引强度(如图 2 所示：)。
- 如果想得到其中两位作者的同被引信息，则将另外一位作者加入“Additional Authors”，然后提交系统处理。如果进一步得到引证文献，则可点击“Go get it”，系统将显示两位作者同时被引的文献题名。如图 3 所示。
- 此系统同时支持 3 位作者的同被引检索。

3.2 AuthorLink 的构成

- 基于 HTML 的检索界面：用于将用户输入的形式转化成 ISI 数据形式，然后将提问输送到 BRS/Search(一种商用搜索引擎)。具体地说，系统将所有类似 Herbert A. Simon, Simon, Herbert A., Simon, HA 等输入形式，自动转换成 SIMON-HA 的提问形式。
- ACA 程序：包括与 BRS/Search 相互影响的一系列C语言程序。当输入一位作者名时，系统产生

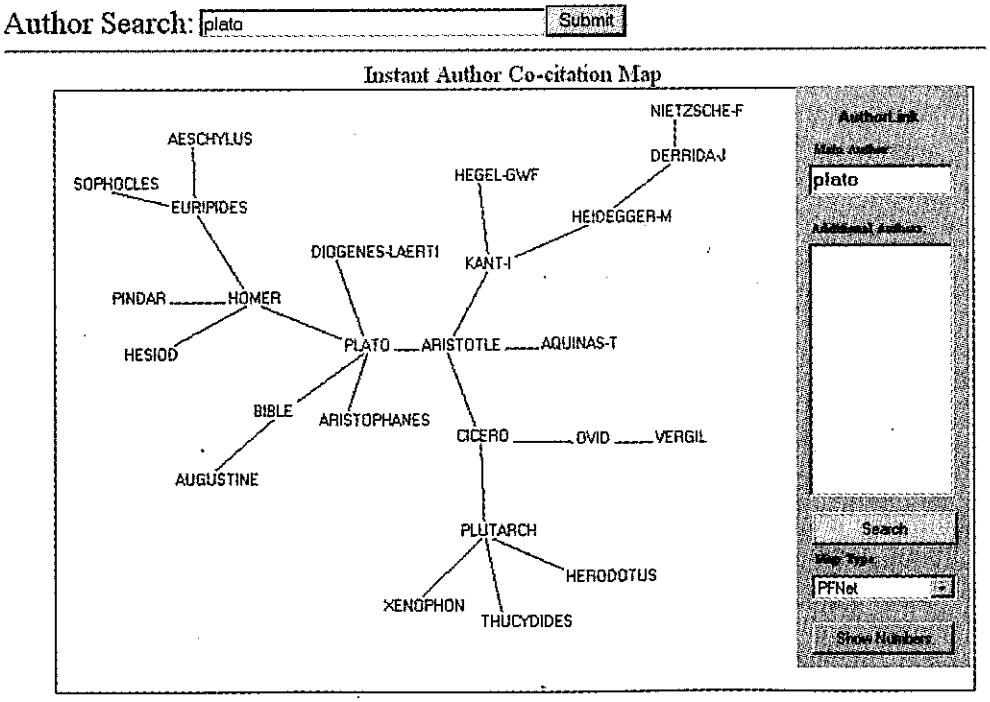

图 2 实时同被引作者图

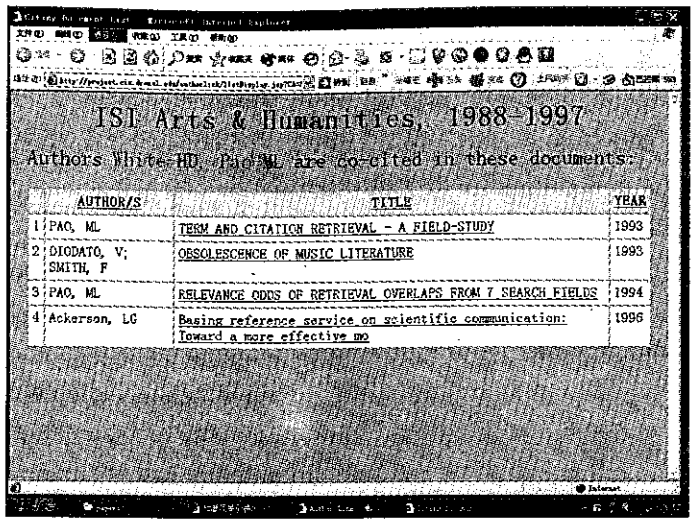

图3 同被引作者的引证文献表

与其高频同被引的24位作者，并利用BRS的TALLY命令排序输出。当系统接到绘图指令时，25×24÷2=300个BRS提问自动输送到BRS，用以获得每一个不重复作者对的同被引频次，所得作者同被引矩阵用于绘图。

● 绘图程序：用于输入同被引矩阵产生绘图结果，并通过Java applet界面显示。目前，系统中有两个绘图程序：自组织映射技术与网络寻址定标法。由于PFNETs对于网络的简化、计算的快捷，而被用于AuthorLink。SOM是典型的神经网络，其中相似的数据逐渐相互靠拢，形成语义相关。每位作者用点代表，作者关系及其聚类用地理区位和相近性代表。同一区位的作者高频同引。两种算法减少了输入数据的复杂性而保持了数据间的关系。

● 交互性图形界面。

3.3 AuthorLink的新突破

传统检索过程将计算机系统中的作者索引与指定的字符串相匹配，而输出单一作者的检索结果，如果想进一步了解与该作者相关的信息必须进行大量额外的工作，阅读大量相关文献。AuthorLink的出现，首先它借助于可视化技术，将复杂的统计结果，用通俗易懂的图像形式显示给用户。此外，它用2-D或3-D的图形代替1-D的线形输出，多角度多层次地反映作者之间的相互关系，几乎相关作者的信息都能在屏幕中同时显示出来，提供整体浏览，由此可以更深入了解作者所从事的研究主题及某一主题的概貌。因此，有人称之为“基于信息计量分析的知识查询系统”，并认为它实现了知识信息提供服务的一次革命。综合起来，AuthorLink具有下述应用：

● 作者的相互链接，有助于用户理解作者的研究范式及主题。

● 学科结构图可用于主题检索，并有助于提高查全率。

● 用户可直观地观察作者的学术关系，甚至挖掘意想不到的关系。

● 仅以一个已知作者作为出发点，即可在“知识地图”中探索不熟悉的领域。

● 能够区分相似的作者。譬如ISI只提供作者的姓和名字的首字母，所以很难通过姓名区分同名作者。AuthorLink则可通过与其联接的作者的学术范围予以区分，从而有助于提高系统的查准率。

4 ACA的未来

信息可视化的发展，使传统ACA添上了翅膀。诚然，AuthorLink仅仅是一种尝试，仍有待于进一步完善。诸如：用户认知与可视图的拟合优度问题，Jan Buzydlowski博士在其学位论文中正在进行相关检验，并将结果用于改善该系统；其次是能否将该系统推广到整个ISI数据库。

此外，ACA自身还有一些细节问题未能解决。首先是数据的收集，到目前为止，仍旧是一件繁琐而费时的工作，并且还需要转化成统计工具或可视化工具所需要的形式。当然，新加坡南洋技术学院最近尝试建立一种网络引文数据库（Web citation database），该系统是用于存储网络出版物引文索引的数据仓储，可直接从网络引文数据库中挖掘作者

信息。其次是作者集团的命名问题。1986年，White曾进行过有关实验，即在统计作者同被引强度的同时，统计被引证文献中同时出现的词的频次，所得词根用于作者集团的命名，这种方法是否科学还有待进一步验证。第三，也是至关重要的问题，即作者相似性计算的优化问题，目前仍在探索中。

无疑ACA作为一种方法在描述科学结构方面是卓有成效的，并且在主题检索方面将大有用武之地。随着信息技术的发展和人们探索的不断深入，ACA会更加成熟。

参 考 文 献

1 H.D.White. Pathfinder networks and author co-citaton analysis: A remapping of paradigmatic information scientists. Journal of the American Society for Information Science and Technology, 2003, 54(5): 423 ~ 434

2 H.D.White. Visualizing a discipline: An author co-citation ananlysis of information science, 1972 ~ 1995. Journal of the American Society for Information Science, 1998, 49(4): 327 ~ 355

3 H.D.White. Author co-citation: A literature measure of intellectual structure. Journal of the American Society for Information Science, 1981, 32(3): 163 ~ 169

4 H.D.White. Author co-citation analysis and Pearson's r. Journal of the American Society for Information Science and Technology, 2003, 54(13): 1 250 ~ 1 259

5 Xia Lin, H. D. White, Jan Buzydlowski. Real-time author co-citation mapping for online searching. Information Processing and Management, 2003, 39(5): 689 ~ 706

6 Yulan He, Siu Cheng Hui. Mining a web citation database for author co-citation analysis. Information Processing and Management, 2002, 38(4): 491 ~ 508

7 罗龙艳. 基于可视化技术的信息检索探讨. 现代图书情报技术, 2002, (4): 36 ~ 38

（责任编辑 许增棋）