科研团队竞争力可视对比分析方法

王杨[1,2)], 余敏楮[1,2)], 单桂华[1)*], 陈恺心[1,2)], 安逸菲[1,2)], 陆忠华[1)]

[1)] (中国科学院计算机网络信息中心先进交互式技术与应用实验室 北京 100190)
[2)] (中国科学院大学 北京 100049)
(sgh@sccas.cn)

摘 要: 学术文献数据的爆炸增长使得找出重要的科研团队变得十分困难. 如何评估团队的科研竞争力, 比较不同的科研团队的科研成果, 发现核心团队等问题变得越来越重要. 为此, 基于文献数据, 构建了科研团队竞争力分级指标体系; 通过引入派系强度改进了派系过滤算法, 用于寻找合作网络中的科研团队; 基于科研团队竞争力分级指标体系, 设计了可视化对比分析方法, 用于对比同一领域中不同的科研团队的科研竞争力. 以天文领域真实的科研团队数据为例, 实验结果和专家验证, 证实了方法的实用性和有效性.

关键词: 文献数据; 科研竞争力; 可视分析; 合作网络
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2020.18174

Visual Comparison Analysis of the Competitiveness of Scientific Research Groups

Wang Yang[1,2)], Yu Minzhu[1,2)], Shan Guihua[1)*], Chen Kaixin[1,2)], An Yifei[1,2)], and Lu Zhonghua[1)]

[1)] (*Advanced Interactive Technology and Application Laboratory, Computer Network Information Center, Chinese Academy of Sciences, Beijing* 100190)
[2)] (*University of Chinese Academy of Sciences, Beijing* 100049)

Abstract: The explosive growth of academic publication data makes it very difficult to study research groups. How to evaluate and compare the competitiveness of research teams, and identify core groups is becoming more and more important. By leveraging the publication data, this paper proposes a hierarchical index system for depicting the competitiveness of research groups. We introduce an enhanced clique percolation method (CPM) by using the clique-strength to find research groups in co-author network. We design and implement a visual comparison approach to support interactive exploration of the competitiveness of different research groups in a field. Experiments and expert reviews on the case in the field of astronomy demonstrate the usefulness and effectiveness of our approach.

Key words: publication data; research competitiveness; visual analysis; co-author network

当今世界上每年都有大量的学术产出, 学术数据正在迅速增加. 到目前为止, 出版物超过 2.3 亿, 作者超过 2.3 亿, 研究主题超过 70 万, 学术会议数量超过 4 千, 学术期刊超过 4.8 万, 研究机构

收稿日期: 2019-12-09; 修回日期: 2019-12-16. 基金项目: 中国科学院“十三五”信息化专项课题(XXH13504). 王杨(1981—), 男, 硕士, 工程师, CCF 会员, 主要研究方向为信息可视化、可视分析、人工智能; 余敏楮(1990—), 女, 硕士, 工程师, CCF 会员, 主要研究方向为可视分析; 单桂华(1976—), 女, 博士, 研究员, 硕士生导师, CCF 会员, 论文通讯作者, 主要研究方向为信息可视化、可视分析、科学可视化、智能交互、虚拟现实; 陈恺心(1996—), 女, 硕士研究生, 主要研究方向为信息可视化、可视分析; 安逸菲(1996—), 女, 硕士研究生, 主要研究方向为科学可视化、可视分析; 陆忠华(1965—), 女, 博士, 研究员, 博士生导师, CCF 会员, 主要研究方向为高性能计算、网格应用.

超过 2.5 万. 同时，随着信息网络的日渐成熟，不同学科、不同机构之间的科研合作越来越紧密. 科研合作已经成为促进高质量科研成果产出与推进科研创新的强大动力. 目前，如何从海量的学术数据中理清合作关系，找出特定领域的核心科研人员和核心科研团队已经成为情报学、信息科学等学科领域的研究热点. 因为核心科研人员和核心科研团队是深入分析学科发展态势、发现领域前沿的基础.

此外，挖掘和分析特定领域的核心科研人员和核心科研团队是项目资助者、科技政策制定者决策前的必要环节. 相关工作人员不仅需要了解特定领域有哪些专家，有哪些核心科研团队，还需要从多个角度全方位地分析专家和科研团队的优势和不足. 除了团队论文的数量，还需要看其质量以及在各细分领域的表现情况，各团队成员的贡献情况，并深入分析专家学者的合作与交流是否在一定程度上促进了科研成果的产出，投入和产出是否相关，合作交流是否有特定的规律等. 其次，团队在所属领域的影响力往往是科研团队实力的表现. 因此，从多个角度综合制定竞争力指标是分析研究团队竞争力重要且必要的工作. 通过分析这些指标，可以帮助投资者确定哪个团队更能胜任特定任务，帮助科技政策制定者确定哪些不足需要通过政策来指导.

为了找出特定领域最具潜力的科研团队和最佳资助团队，了解不同团队的在该领域科研实力优劣，进行多维度对比分析是必不可少的环节. 通过在多个维度进行对比分析，相关人员可以结合经验来判断不同科研团队的竞争优势. 然而，为了较全面地对比分析 2 个团队，需分析的维度及对应的统计数值会非常多，这大大增加了对比分析的难度. 当需要深入分析在某个 A 团队擅长的领域中 B 团队的表现如何时，如果用列表的形式呈现数据，那么需要呈现 2 个团队在各个领域的论文数量情况、论文质量情况等信息，数据量将会非常大，查找将会极其不便. 针对上述不直观、不高效等分析难点，可以通过可视化技术将数值的大小映射到图形的长度等视觉通道来直观地表达数值的大小及其程度，通过交互技术提供良好宏观/微观洞察来高效地分析数据.

本文提出了一个指标体系用于综合衡量科研团队竞争力，并改进了派系过滤算法(clique percolation method, CPM)用于从论文数据中挖掘科研团队；然后针对提出的指标体系设计了一套可视分析方法，为更好地对比分析 2 个科研团队提供直观、友好、高效的交互界面. 最后，本文基于学术数据库，对某一领域的论文数据进行了清洗、处理和分析，并按照一定标准选择了 2 个较为典型的专家团队进行可视分析，以验证本文方法的有效性.

1　研究现状

国内外竞争力分析的相关工作主要集中在对领域、机构的整体竞争力进行分析，对更细粒度的科研团队的竞争力分析较少. 此外，对科研团队的分析主要集中在团队绩效分析与评估、团队的组织和建设方面，科研团队的竞争力分析仅作为团队绩效分析的一部分进行研究. 在科研团队竞争力指标体系方面，与本文最相关的工作为王衍喜[1]基于统计学中指标选取的系统性原则、科学性原则、可比性原则和可行性原则提出的学科团队竞争力评价指标体系. 该指标体系主要包括了多种类型论文的数量、硕士人数、博士人数、团队人数、研究员人数和副研究员人数. 此外，Keathley-Herring 等[2]通过研究 1983—2016 年期间发表的 123 篇领域成熟度评估相关文章，提出了一个通用的研究领域成熟度评估指标体系，包括专利、论文、项目、研究人员和影响力等多个维度的指标.

Gleicher 等[3-4]总结了对比可视分析技术关键技术，将其分为 3 大类：并置(juxtaposition)、叠加(superposition)和显式编码(explicit encoding). 并置是指将待比较对象并排放置，分析人员通过来回查看对象相对应的数据进行比较. 叠加是指将待比较的对象重叠放置在同一参考系下，分析人员通过观察不同对象在同一参考系上对应位置的区别进行比较. 显式编码是指直接呈现对象之间的区别，无需分析人员进行计算. 通常，对比可视分析采用其中一种或多种技术. 在竞争力对比可视分析应用方面，Henry 等[5]采用密集子图的方法来对合作网络进行凝聚子群分析. Chinchilla-Rodríguez 等[6]则提出了通过不同的维度来展现合作关系，如在机构之间、国际之间的合作. Vuillemot 等[7-8]将作者排名表与折线趋势图创造性的结合，使读者可以对排名的变化趋势与作者信息有一个总体的概览，并很好地实现了与用户的交互. Görg 等[9]提供了一个通过列表视图显示作者的元数据，如共同作者或与他有关的概念的集成可视化.

在科研团队挖掘方面，Newman[10-11]在对网络进行社团划分聚类的算法中定义了模块度来衡量划分结果的优劣．此后引出了一系列基于模块度对社团划分进行优化的方法：Palla 等[12]于 2005 年在《Nature》上提出了派系过滤算法 CPM；Gupta 等[13]提出的 ENetClus 算法关注随着时间推移异构网络所产生的变化，其研究社团的产生、发展或消退和社团之间的互相影响．国内对科研合作关系的研究也一直在推进，大多数关注在理论阐述方面，或从网络特性方面进行计量分析，算法方面的研究较少．其中，张鹏等[14]利用层次聚类法来对研究人员的合作网络进行聚类分析；梁艳琪等[15]则采用了复杂网络以及社会网络的分析方法来对科研社会网络进行可视化与聚类分析．

2 科研团队竞争力指标体系

本文旨在建立一个基于文献数据的科研团队竞争力指标体系．通过分析国内外在竞争力指标体系、成熟度指标体系等方面的相关工作，本文归纳总结并筛选出其中适合作为科研团队竞争力分析指标的部分．由于科研合作对科研产出的促进作用与日俱增，本文在指标体系中增加科研合作方面的指标，便于对比分析不同科研团队的合作模式．本文提出的基于文献数据的科研团队竞争力指标体系如表 1 所示．

表 1 科研团队竞争力指标体系

一级指标	二级指标	示例
科研产出	论文数量	各类期刊论文数，影响因子分布…
	论文质量	篇均合作机构数，篇均合作作者数…
科研合作	合作深度	篇均合作机构数，篇均合作作者数…
	合作广度	合作机构数，合作作者数…
科研地位	论文影响力	被引次数，篇均被引频次…
	团队领头人影响力	课题组长 H 指数…

该指标体系主要包括 3 个一级指标和 6 个二级指标．一级指标有科研产出、科研合作和科研地位．

科研产出下设 2 个二级指标，分别为论文数量、论文质量．论文数量指标通常可以通过历年论文数、历年 SCI 论文数、论文总数、SCI 论文总数、人均论文数、人均 SCI 论文数、高质量论文数、人均高质量论文数、各领域论文数量及历年各领域论文数量等具体数据来体现．论文质量指标通常可以通过论文的期刊分布、论文所投期刊影响因子分布等具体数据来体现．

科研合作下设合作深度、合作广度 2 个二级指标．合作深度指标通常可以通过合作作者深度、合作机构深度等具体数值来体现．合作作者深度是指某科研团队论文的篇均合作作者数，合作机构深度是指某科研团队论文的篇均合作机构数．合作广度指标通常可以通过合作作者广度、合作机构广度等具体数值来体现．合作作者广度是指合作的作者数，合作机构广度是指合作的机构数．

科研地位下设论文影响力、团队领头人影响力 2 个二级指标．论文影响力通常通过论文的总被引次数、篇均被引次数等具体数据来体现．团队领头人影响力通常可以使用 H 指数等来体现．

3 基于改进的 CPM 的专家团队挖掘方法

分析科研团队竞争力需要先找出领域内的所有科研团队，再选择目标对象进行分析．本文基于改进的 CPM 来发现科研网络中的科研团队．

3.1 CPM 的基本原理

在 CPM 中，一个典型的社区是几个具有共享节点的完全子图的集合，这样的定义可以满足实际情况中社区内成员连接紧密而社区间连接稀疏的要求．定义一个 k-派系为一个具有 k 个节点的完全子图，如图 1 所示，当 2 个 k-派系之间共享 $k-1$ 个节点时，则认为这 2 个 k-派系连通，而一个 k-派系社区是所有互相连通的 k-派系的集合．这样，在一个 k-派系社区中，社区中的成员会与很多同一社区内其他成员相连，但并不一定与每一个节点都连接；并且节点可以属于不同的社区，形成了社区间的重叠性，如图 2 所示．算法的目的在于找到社会网络中这样的 k-派系社团．

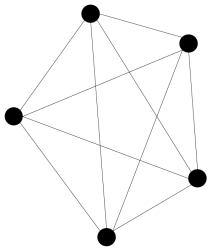

图 1 5-派系社团

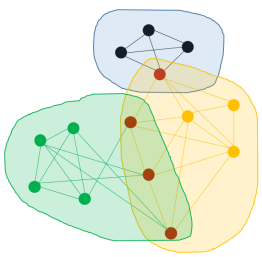

图 2 社团之间的重叠性

首先，算法会找到网络中所有的大小不同的派系(极大完全子图)，然后计算派系之间共享的节点数目，建立派系的重叠矩阵；之后对于给定的参数 k，可以得到 k -派系的邻接矩阵，这样连通的部分就可以构成一个 k -派系社团．算法的具体步骤如图 3 所示．

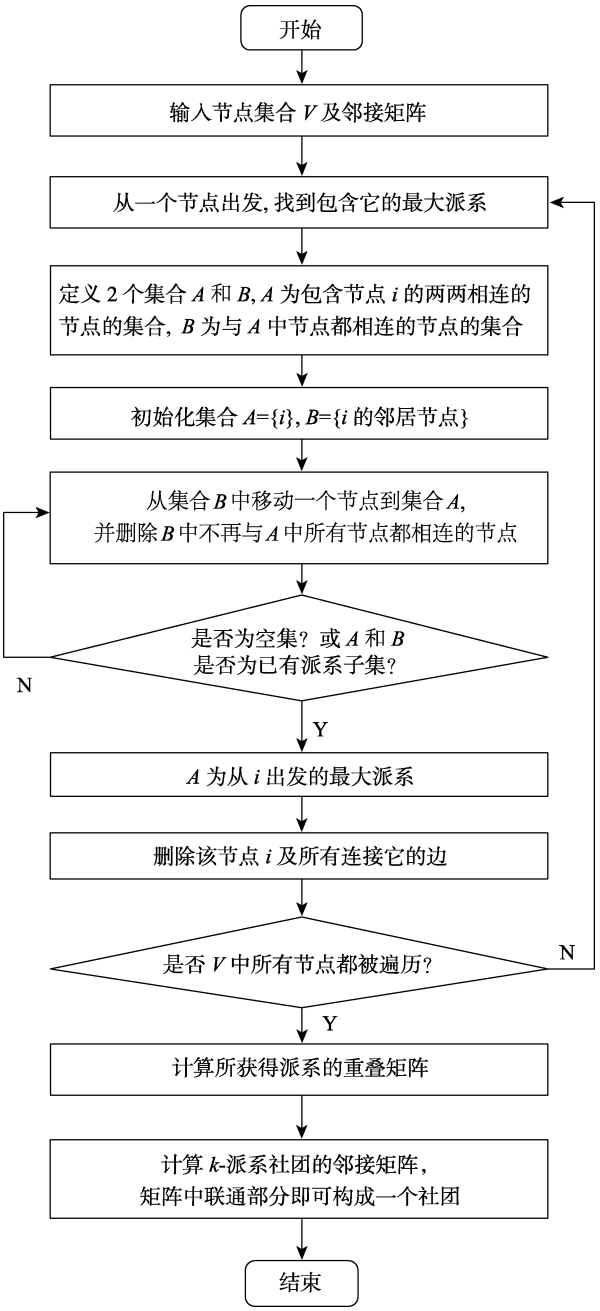

图 3　CPM 流程图

3.2　改进的 CPM

事实上，CPM 只适用于简单的无向无权图，对有权图的情况，算法会通过设定一个阈值，并只保留权值大于该阈值的边的方法，将其转化为无权图．

而对于科研合作网络来说，边权值都为离散的整数型变量，最低的权值为 1，即 2 人之间有过一次合作，否则 2 人之间没有连接边．若将阈值设为大于 1 的数值，会导致很多权值为 1 的边被过滤掉，进而很多合作情况被忽略．这样在生成社区的过程中，一些本来可以构成完全子图的结构会被遗漏，影响了结果的准确性．若将阈值设为小于 1 的数值，则不会起到任何筛选的作用，网络会被当作无权图处理．

因此，针对本文所要挖掘的科研合作网络，需要对算法做相对应的改进．

对于无向有权图的情况，需要考虑边的权值对派系寻找和在形成社团的过程中所产生的影响．在这里，可以引入对于派系(子图)强度的定义，其定义为其边权值的几何平均数[16]，即

$$I(c)=\left(\prod_{\substack{i<j\\ i,j\in c}}\omega_{ij}\right)^{\frac{2}{n(n-1)}}.$$

其中，对于一个含有 n 个节点的完全子图 c 来说，它有 $\dfrac{2}{n(n-1)}$ 条边；ω_{ij} 即为节点 (i,j) 之间连接边的权值．

建立对于派系强度的定义后，就可以对算法中所要寻找的派系进行过滤．为此可给定阈值，并在后续构建社团时只考虑强度高于此阈值的派系．

这样，既可以在挖掘团队的过程中对合作网络中的一些噪声扰动完成筛选和过滤，更加精准地建立作者的核心团队；同时又不会过滤掉大量的连接边，丢失重要的关键信息．例如，当设 $\omega_{ij}>1$ 时，很多边权值为 1 的连接边仍然会被保留．这样的定义很好地考虑了 2 个强连接的合作团队之间可能存在的弱连接情况，使得这样的弱连接在构建社团时不会被移除．

3.3　团队挖掘结果评估方法

对专家合作团队的挖掘是用来在实际的合作网络中发掘未知的社区，因此如何在没有准确结果的情况下评价算法所发现的社团的质量尤为重要．模块度作为一种衡量社区结构质量的标准被广泛使用，它定义为落在社区内的连接边的比例减去一个随机期望，即当网络对边进行随机分配时所得到的比例

$$Q=\frac{1}{2m}\sum_{ij}\left(A_{ij}-\frac{k_ik_j}{2m}\right)\delta(c_i,c_j).$$

其中，m 为网络中边数；A_{ij} 表示节点 i 和 j 之间的

连接状态，定义为

$$A_{ij}=\begin{cases}1, & \text{如果 } i \text{ 与 } j \text{ 相连}\\ 0, & \text{其他}\end{cases}.$$

用 k_i 表示点 i 的度，即 $k_i=\sum_j A_{ij}$；$\delta(c_i,c_j)$ 则用来判断节点 i 和 j 是否属于同一个社区，若是，则 $\delta(c_i,c_j)=1$，否则 $\delta(c_i,c_j)=0$.

该方法可以有效地衡量社区结构的划分情况，但是其无法适用于可重叠的社区结构，同时也没有考虑边权值的影响.

因此，本文引入一种可以用于重叠结构的度量方法[17]. 该方法认为，一个好的社区划分应该满足 2 点：(1) 社区内某一点 i 应该在内部连接较多与外部的连接边较少；(2) 社区内的节点分布应该是较为密集的. 基于这样的认识，该度量可用公式表示为

$$M^{\mathrm{ov}}=\frac{1}{k}\sum_{r=1}^{k}\frac{\sum_{i\in c_r}\dfrac{\sum_{j\in c_r,i\neq j}a_{ij}-\sum_{j\notin c_r}a_{ij}}{d_i\cdot s_i}}{n_{c_r}}\cdot\frac{n_{c_r}^{\mathrm{e}}}{C_{n_{c_r}}^2}.$$

其中，$\dfrac{n_{c_r}^{\mathrm{e}}}{C_{n_{c_r}}^2}$ 定义为第 r 个团队 c_r 的密度，$n_{c_r}^{\mathrm{e}}$ 为它的边数，$C_{n_{c_r}}^2$ 为它是完全子图的边数；$\sum_{j\in c_r,i\neq j}a_{ij}$ 表示点 i 与团队内部点 j 的边权值总和，$\sum_{j\notin c_r}a_{ij}$ 表示点 i 与团队外点 j 的边权值总和；d_i 为点 i 的度，而 s_i 为点 i 所属于的团队的个数；n_{c_r} 为 c_r 所包含的点的个数. 在此基础上，本文综合了以上度量标准并进行了优化. 为了防止算法将网络划分为少数几个极大的社区，需要对社区的个数进行约束，即认为一个好的划分所得到的社区个数应尽可能多，但又不会形成一个巨大的社区使得社区结构特点被抹掉；同时，考虑未被聚类的节点，即那些不属于任何一个社区的节点，应该尽可能少；并且，为了更好地展现社区结构划分的质量，应将评价得分尽可能离散的分布在(0, 1)区间内，越接近 1，代表划分结果越好. 由此，得到最终的评价标准为

$$M=\frac{Q+M^{\mathrm{ov}}}{2}\times\frac{k}{n_{\mathrm{u}}}.$$

其中，n_{u} 表示网络中未被聚类的节点的个数.

4 科研团队竞争力可视化对比分析

4.1 需求分析

根据第 3 节提出的科研团队竞争力指标体系，科研团队竞争力对比分析需要从宏观和微观 2 个角度去分析. 在宏观角度，需要能展示指标体系中科研产出(论文数量、论文质量)、科研合作(合作宽度、合作广度)、科研地位(论文影响力、团队领头人影响力) 3 大维度的概况，既要能展示单个团队在各个维度上的概况，又要便于对比 2 个团队在各个维度上的差别. 在微观角度，要能展示科研团队在二级指标下的各个具体指标的数值. 由于指标较多，2 个团队的数据很难在有限的屏幕上一次性展现，尤其是 2 个团队在各个领域的论文数量及其对比情况、在各个期刊上发表的论文数量及其对比情况. 同时，数据维度的繁多大大增加了大脑记忆负担，使微观层面对比难上加难. 因此，需要针对需求设计相应的可视化方法，以交互方式合理地增加信息量，并分层次地呈现数据. 本文归纳需求如下：

R1. 需要一个概览图简要概括 2 个团队人数、论文的数量和质量、合作情况以及科研影响力；

R2. 能清晰地呈现 2 个团队的论文数量以及数量上相差的程度；

R3. 能清晰地展现 2 个团队各个成员的贡献度，便于发现团队成员的科研实力情况，进而对比 2 个团队结构；

R4. 能清晰地对比 2 个团队在各个研究热点上的论文数量，从而快速获知 2 个团队的兴趣点的异同，以及谁是某一研究热点的主要贡献者；

R5. 能清晰地展现 2 个团队的期刊选择偏好，以及在不同领域不同影响力的期刊上发表论文的情况，便于发现团队是否具有核心竞争力，核心竞争力的区别在哪里；

R6. 能清晰地呈现 2 个团队的合作情况，包括合作机构、合作作者，以对比它们的合作模式的异同，分析不同的合作模式对科研成果的影响.

4.2 方案设计

针对上述需求，本文设计了可视化对比分析方法用于对比 2 个科研团队的科研竞争力，包括图 4 的团队概要信息，图 5~图 7 的论文数量对比分析，图 8 的论文质量对比分析，图 9 的科研合作对比分析 4 大部分.

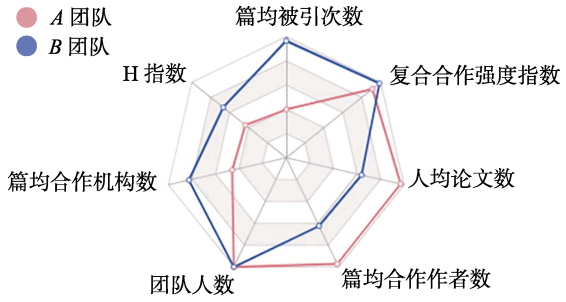

图 4　团队概要信息

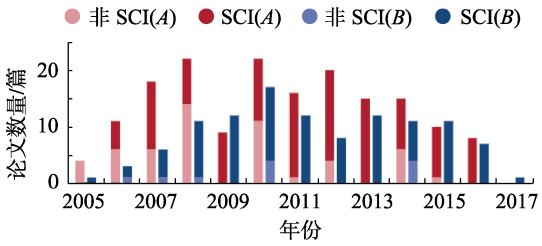

图 5　历年论文数量对比分析

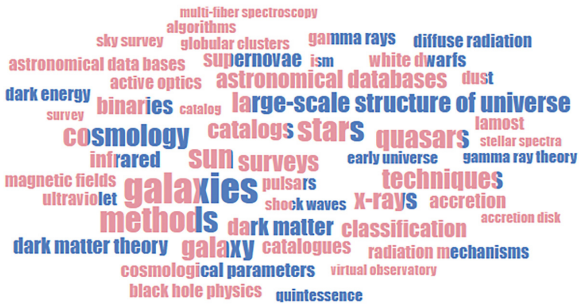

图 6　各关键词论文数量对比分析

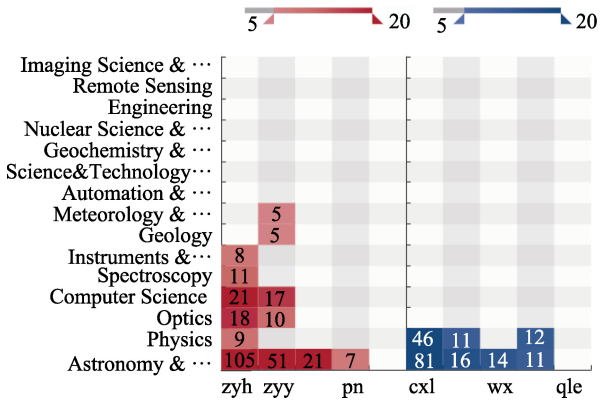

图 7　各领域论文数量对比分析

团队概要信息采用雷达图的形式从 7 个维度展现和比较科研团队的学术状况. 采用雷达图将 2 个团队各个维度的数据放置在同一个极坐标系中, 用极径的长度(长度是人眼最敏感的视觉映射通道之一)来表达数值的大小, 不仅能清晰地展现它们在各个维度上的表现, 更易于对比(R1). 图 4 中的

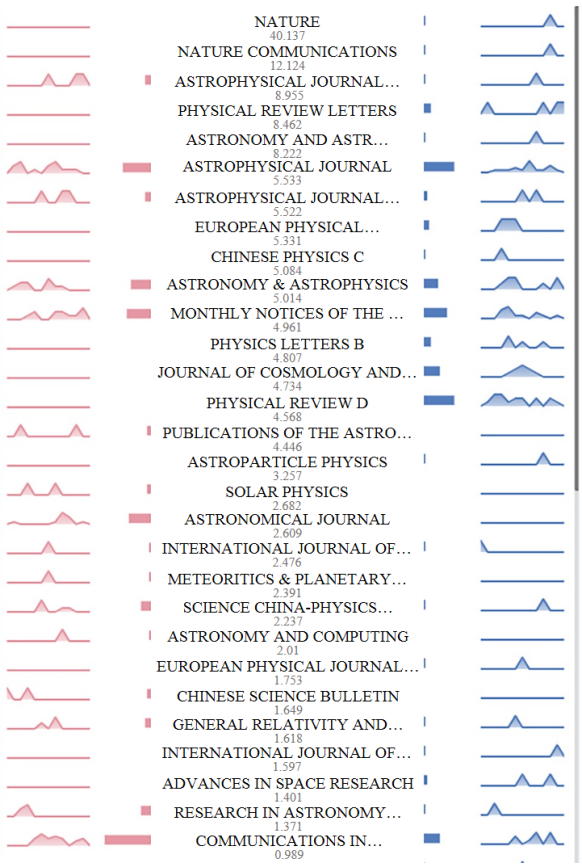

图 8　论文质量对比分析

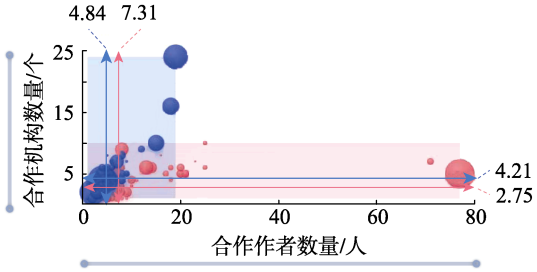

图 9　科研合作对比分析

7 个维度是从科研团队竞争力指标体系中筛选出来的具有简要概括能力的指标, 用于为用户提供直观全面的对比分析功能. 7 个维度分别是人均论文数、团队人数、篇均合作机构数、篇均合作作者数、复合合作强度指数、篇均被引频次及核心专家 H 指数. 其中, 复合合作强度是基于合作作者深度和合作机构深度的复合指标, 其定义如下.

复合合作强度指数(compound collaboration strength, CCS)[18]用于评价科学家合作水平的复合合作强度, 不仅考虑合作的数量, 还考虑合作作者、合作机构的数量和平均值、合作文章的分布均度及合作论文数所占比例等因素, 综合反映科学家的合作行为. 因此, 将该指数直接应用于团队分

析中，复合合作强度指数为

$$\mathrm{CCS}=\sqrt{\ln c_{\mathrm{wa}}\times\ln c_{\mathrm{wi}}\times c_{\mathrm{sa}}\times\frac{c_{\mathrm{si}}}{c_{\mathrm{e}}}\times c_{\mathrm{d}}}\ .$$

其中，c_{wa} 是合作作者宽度，指某科研团队拥有的合作作者数量；c_{wi} 是合作机构宽度，指某科研团队拥有的合作机构数量；$c_{\mathrm{sa}}=N_{\mathrm{a}}/N$ 是合作作者深度，指某科研团队论文的篇均合作作者数，其中 N_{a} 为合作者累计论文数，N 为 SCI 论文总数；$c_{\mathrm{si}}=N_{\mathrm{i}}/N$ 是合作机构深度，指某科研团队论文的篇均合作机构数，其中 N_{i} 为合作机构累计论文数；c_{e} 是合作机构均度，用方差衡量某科研团队所有论文的合作机构数量是否均匀分布，为某科研团队每篇论文所拥有的合作机构数量的 STDEV.S 标准差；$c_{\mathrm{d}}=C/N$ 是合作机构密度，指拥有合作机构的文章数量占该科研团队所有文章总数的比例，其中 C 为与其他机构合作论文数.

在团队概要信息部分，不同的颜色对应不同的团队(红色表示 A 团队，蓝色表示 B 团队)；其他图表中的颜色映射与团队概要信息部分一致. 可以通过图例选择一个团队，也可以选择 2 个团队，同时其他图的数据也同步更新成所选团队的数据. 当鼠标上移到红色线条或蓝色线条上，会出现浮动窗口提示相应团队在各维度上的详细数据.

本文从 3 个方面对比分析论文数量，分别为图 5 所示的历年论文数量对比，图 6 所示的各关键词论文数量对比，以及图 7 所示的各领域论文数量对比.

历年发表论文的数量、历年 SCI 论文的数量的展现融合了并置和叠加 2 种技术，如图 5 所示. 基于堆积柱状图展示了 2 个团队历年发表论文的数量、历年 SCI 论文的数量(R2). 横坐标表示年份，纵坐标代表论文数量，不同系列的颜色代表了不同团队，其中饱和度较高的颜色代表 SCI 论文，用于突出呈现高质量论文的数量. 通过点击图例，可以单独查看一个团队论文/SCI 论文数据.

图 6 对词云进行了改进，每一个关键词拥有 2 种不同比例的颜色，颜色的比例代表 2 个团队发表了拥有该关键词的论文数量的占比. 关键词的大小表示 2 个团队发表了拥有该关键词的论文总数. 为了能深入分析 2 个团队在各个研究热点上的表现，此处设计了鼠标点击交互. 通过选择关键词，图 5 和图 7 更新成所选关键词对应的数据(R4).

图 7 是基于热力图设计的领域论文数量对比图，用于对比 2 个团队的各成员在各个研究热点、研究领域(一个研究领域包含多个研究热点)发表论文的分布情况(R3，R4). 横坐标代表 2 个团队中的不同专家，纵坐标代表研究领域，如果专家在某领域中发表过论文，那么就用颜色填充相应坐标的方格，发表论文数量越多，颜色越深，论文数量的详细数字也显示在方格内. 顶部的数据选择器可以用于选择论文的数量范围. 当用户选择一个展示范围后，该图只显示满足所选范围的方格，便于用户集中注意力分析感兴趣的部分.

论文质量对比分析图如图 8 所示，其将 2 个团队在各类期刊上的历年发表论文数量曲线图、历年论文总数条形图分别放置在中轴线两侧(采用了并置技术)，并将期刊按照影响因子从高到低排序(采用了叠加技术)，能直观、高效地对比 2 个团队发表论文的质量情况和期刊偏好(R5). 鼠标移至某一期刊，该期刊的信息会被放大便于查看；点击某一期刊，图 5~图 7 和图 9 更新成与所选期刊相关的数据.

科研合作对比分析基于散点图进行改进，采用了叠加的方法，如图 9 所示在同一笛卡儿坐标系中展现 2 个团队的合作信息，便于对比 2 个团队的合作特点(R6). 横坐标表示合作者数量，而纵坐标表示合作机构数量. 每个点代表一篇论文，不同颜色表示不同科研团队，每个点的大小与论文引用次数正相关，半透明颜色的矩形定义了对应颜色科研团队的合作范围. 矩形的长度是合作者的宽度，矩形的宽度是合作机构宽度. 水平双向箭头标记表示合作机构深度，即平均合作机构数. 垂直双向箭头标记表示合作作者深度，即平均合作者数. 通过左侧和底部的数据选择器，可以选择任意范围内的数据，便于查看数据密集区域的信息.

5 案例分析

本文选择了 2 个较为典型的团队(本文简称 A 团队和 B 团队)作为案例研究，他们是同一研究领域中具有相同团队规模和相近论文数量的科研团队. 本文默认一个团队中论文数量最多学者是该团队的核心专家. 出于隐私原因，文中隐去了真实姓名.

图 4 显示了 2 个团队的学术竞争力概要，红色表示 A 团队，蓝色表示 B 团队. 从图中可以看出 A 团队的人均产出大于 B 团队，且能明显看出团队的人均产出比 B 团队多大约 2/3. 但 B 团队的科研影响力(核心专家 H 指数，篇均被引频次)远大于

A 团队．*A* 团队倾向与少量机构的多个学者合著文章，而 *B* 团队倾向与遍布多个机构的少量学者合著文章.

从图 5 中可以看出，除了 2009, 2013, 2016 年，2 个团队发表的论文均为 SCI 论文之外，*A* 团队每年的论文总数较多，SCI 论文数量较多但占总论文数量的比重较低；*B* 团队每年的论文总数较少, SCI 论文数量较少但占总论文数量的比重较高. 由此可见，*A* 团队在论文总数、SCI 论文数量上超过了 *B* 团队，但 *B* 团队的高质量论文占比高于 *A* 团队.

从图 6 可以看出，*A* 团队和 *B* 团队主要研究“galaxies, supernovae, radiation mechanisms, cosmological parameters”等，且 2 个团队在这些研究点上势均力敌．此外 *A* 团队更侧重于“methods, active optics, sun, stars, x-rays, classification, astronomical databases”等研究点，而 *B* 团队更侧重于“cosmology, large-scale structure of universe, diffuse radiation, dark matter theory”等研究点.

在各领域论文数量对比图中，图 7 所示为筛选论文数大于 5 的研究领域，通过该图可以看到 *A* 团队研究领域较宽泛，而 *B* 团队多集中在 2 个领域. 此外，可以发现，*A* 团队中的 2 位学者 zyh 和 zyy 在该团队中是主要贡献者，zyh 的贡献尤其突出；而 *B* 团队中，cxl 的实力远远超过了团队其他学者的实力.

从图 8 可以看出，*A* 团队的论文分布在各个期刊上，在影响因子较高到较低的期刊均有论文发表；而 *B* 团队的论文主要分布在影响因子较高的期刊上，在影响因子较低的期刊上文章数量非常少. 值得注意的是，*B* 团队近几年在《Nature》上发表了一篇论文.

从图 9 可以看出，红色矩形和蓝色矩形面积相近，说明 2 个团队的合作范围相似，但红色矩形较长，蓝色矩形较宽，表明 *A* 团队更偏向与少量机构内的多位学者合作，*B* 团队偏向和不同机构中的个人合作；*B* 团队的合作机构数量高于 *A* 团队，而 *A* 团队的合作者数量高于 *B* 团队. 此外，*A* 团队的高引用论文位于红色矩形的右下方，这意味着共同作者较多时，该团队的论文质量较高，受合作机构数量影响较小；共同作者和合作机构数量对 *B* 团队论文质量似乎没什么影响，因为在蓝色矩形的右上方和左下方均有较大的蓝色散点.

综上所述，虽然 *A* 团队发表的论文数量较多，但 *B* 团队的论文质量和影响力远远高于 *A* 团队．2 个团队的合作侧重点不同，*A* 团队的合作者宽度更大，*B* 团队的合作机构宽度更大，更多的合作有助于 *A* 团队提高合作质量，但论文合作似乎对 *B* 团队的论文质量没有显著影响．*B* 团队的研究范围较窄，较为专注，注重高质量的研究，*A* 团队涉及领域较多，选择期刊也较多，论文数量也远高于 *B* 团队．因此，可以得出 *B* 团队在学术研究中比 *A* 团队更具竞争力的结论.

此外，为了验证分析的有效性，本文咨询了该领域 2 位熟知上述 2 个团队的专家，专家表示分析结果与事实基本一致：研究领域和研究内容与事实相符；*B* 团队的论文质量和影响力确实较高，且在 2015 年在《Nature》期刊上发表了一篇论文. 但是，统计的论文数量比 2 个团队实际的论文数量少几篇. 经过分析发现，主要原因是数据处理过程中剔除了少量数据质量不佳的文章导致的.

6　总结与展望

本文提供了一套可视化分析方法来分析科研团队的学术竞争力，主要成果有以下 3 点：

(1) 构建科研竞争力指标体系. 基于国内外有关竞争力分析的研究进展，借鉴其在指标体系设计中所考虑的相关因素和指标，从科研产出、科研合作、科研地位 3 个方面进行设计. 从多个方面全面准确地对科研团队的竞争力进行描述.

(2) 基于改进的 CPM 挖掘专家团队. 构建专家的科研合作网络，并引入 CPM 对网络中的合作团队进行挖掘. 由于 CPM 只考虑无权图的情况，本文针对有权的合作网络图对 CPM 做了优化改进，加入了边权值，规避了将 CPM 应用到有权图时可能产生的缺陷.

(3) 设计可视化方案. 以可视化的形式展现评估可以表现科研团队学术竞争力的各种数据、信息和趋势，并在不同的团队之间进行对比. 通过系统、有效的竞争力评价，获得团队的专长与劣势，展现不同科研团队的研究状态和研究质量，使学术机构的决策者和投资者全面掌握组织、领域中所有科研团队学术状态和核心竞争力信息，了解科研团队的最新研究动态. 最后，通过对选取的 2 个典型案例团队 *A* 和 *B* 的研究，详细分析并对比了它们在不同维度上的指标，得出 *B* 团队在学术研究中比 *A* 团队更具竞争力的结论；同时验证了本文方法在分析研究团队的学术状况和竞争力方面

的有效性和实用性.

当然, 本文也有一些不足之处. 其中的可视化方案多为传统的可视化图表, 或在传统可视化图表基础上进行改进, 仅满足分析与展示的要求, 未能在图表上有颠覆性的创新. 此外, 本文提出的科研竞争力指标体系仅仅基于论文数据, 由于专利、项目等信息不易获取, 且不同领域具有不同的数据及特征, 其他维度的指标暂时未纳入指标体系. 未来将进一步探索基于多种数据源的学术竞争力评价指标体系, 并设计新颖的可视化方案, 为大型学术机构的决策者提供更准确的决策支持.

参考文献(References):

[1] Wang Yanxi. Analysis and evaluation on disciplinary team competitiveness: an empirical study[D]. Xi'an: Xidian University, 2011(in Chinese)
(王衍喜. 学科团队竞争力分析与评价实证研究[D]. 西安: 西安电子科技大学, 2011)

[2] Keathley-Herring H, van Aken E, Gonzalez-Aleu F, *et al*. Assessing the maturity of a research area: bibliometric review and proposed framework[J]. Scientometrics, 2016, 109(2): 927-951

[3] Gleicher M, Albers D, Walker R, *et al*. Visual comparison for information visualization[J]. Information Visualization, 2011, 10(4): 289-309

[4] Gleicher M. Considerations for visualizing comparison[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(1): 413-423

[5] Henry N, Fekete J D, McGuffin M J. NodeTrix: a hybrid visualization of social networks[J]. IEEE Transactions on Visualization and Computer Graphics, 2007, 13(6): 1302-1309

[6] Chinchilla-Rodríguez Z, Vargas-Quesada B, Hassan-Montero Y, *et al*. New approach to the visualization of international scientific collaboration[J]. Information Visualization, 2010, 9(4): 277-287

[7] Vuillemot R, Perin C. Investigating the direct manipulation of ranking tables for time navigation[C] //Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. New York: ACM Press, 2015: 2703-2706

[8] Perin C, Vuillemot R, Fekete J D. A table!: improving temporal navigation in soccer ranking tables[C] //Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2014: 887-896

[9] Görg C, Liu Z C, Kihm J, *et al*. Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(10): 1646-1663

[10] Newman M E J. The structure of scientific collaboration networks[J]. Proceedings of the National Academy of Sciences, 2001, 98(2): 404-409

[11] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2): Article No.26113

[12] Palla G, Derényi I, Farkas I, *et al*. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043): 814-818

[13] Gupta M, Aggarwal C C, Han J W, *et al*. Evolutionary clustering and analysis of bibliographic networks[C] //Proceedings of International Conference on Advances in Social Networks Analysis and Mining. Los Alamitos: IEEE Computer Society Press, 2011: 25-27

[14] Zhang Peng, Li Menghui, Wu Jinshan, *et al*. The community structure of scientific collaboration network[J]. Complex Systems and Complexity Science, 2005, 2(2): 30-34(in Chinese)
(张鹏, 李梦辉, 吴金闪, 等. 科学家合作网络的聚类分析[J]. 复杂系统与复杂性科学, 2005, 2(2): 30-34)

[15] Liang Yanqi, Peng Bo, Gao Jinsong. Research collaboration network visualization based on JASIST[J]. Journal of Intelligence, 2015, 34(8): 87-91(in Chinese)
(梁艳琪, 彭博, 高劲松. 基于 JASIST 的科研合著网络可视化研究[J]. 情报杂志, 2015, 34(8): 87-91)

[16] Onnela J P, Saramäki J, Kertész J, *et al*. Intensity and coherence of motifs in weighted complex networks[J]. Physical Review E, 2005, 71(6): Article No.65103

[17] Lázár A, Ábel D, Vicsek T. Modularity measure of networks with overlapping communities[J]. EPL(Europhysics Letters), 2010, 90(1): Article No.18001

[18] Chen Yunwei, Deng Yong, Chen Fang, *et al*. Research on construction and application of CCS index[J]. Library and Information Service, 2015, 59(13): 96-103(in Chinese)
(陈云伟, 邓勇, 陈方, 等. 复合合作强度指数构建及应用研究[J]. 图书情报工作, 2015, 59(13): 96-103)