面向科研评价的学术关系可视化方法研究

王有国[1],刘玉琴[2],汪雪锋[1]
(1. 北京理工大学 管理与经济学院,北京 100081;2. 北京印刷学院 绿色印刷包装产业技术研究院,北京 102600)

摘要:针对目前科研评价中可视化技术应用的不足,改进传统合著关系可视化,融合文本关联关系可视化,并新建研究主体年代引证关系可视化,从学术关系构建和可视化角度综合考察被评价对象科研成果的数量、质量和影响力。作为现有评价方法的辅助手段,提出的可视化方法可用于特定领域科研工作者、机构、地区和期刊评价。在阐述了研究背景和过程后,以3D打印技术SCI论文数据为例进行实证应用。

关键词:科研评价;学术关系;可视化
中图分类号:F224.5 文献标识码:A 文章编号:1002-0241(2014)05-0013-06

0 引 言

科研评价是科研管理工作的重要组成部分,是保证科学研究活动顺利进行的基本保障,对促使科研水平提高、科研交流和科研创新都具有重要作用。如何健全科研评价制度,完善评价体系,改进评价方法,已经成为科学学、科学计量学的重要研究内容之一。

为克服目前科研评价中可视化技术应用的不足,丰富现有科研评价方法,针对特定技术领域、技术主题或学科的科研工作者、科研机构、地区和学术期刊评价,本文从学术关系可视化角度研究海量科研信息中潜在的学术关系,利用计算机支撑的、交互的、对抽象数据的可视化表示,来增强人们对这些科研信息的直观认识。从整体上动态关注学术研究者、研究机构、研究内容的进展以及相互之间的关系,作为现有科研评价方法的辅助手段,揭示被评价对象的科研内容或成果的数量、质量和影响力情况,辅助决策参考。

1 研究背景

1.1 研究内容界定

科研评价的过程是一个多学科融合的过程,评价对象包括科研计划、项目、成果、机构、人员、期刊等[1],评价方法因评价对象、内容、标准的差异而不同。常用的科研评价方法包括:同行评议、德尔菲调查等专家评价法;灰色系统、层次分析、模糊评判、数据包络分析等综合评价法;主成分分析、因子分析等统计学方法;引文分析、词频分析等文献计量学方法。本文从学术关系可视化的角度进行科研评价,不在于建立完整、全面的科研评价体系,而是力图将学术关系可视化技术作为现有科研评价方法的辅助手段,发现以往研究方法中被忽视的内容,克服目前科研评价中可视化技术应用的不足。为此,本文研究的科研评价方法限定为基于科技文献数据的、针对特定技术领域、技术主题或学科的科研工作者评价、科研机构、地区和学术期刊评价,是一种基于文献计量学的科研评价。

1.2 文献回顾

伴随着科学技术的迅猛发展,科研难度日趋加大,学科间渗透交叉越来越普遍,学术研究主体之间呈现既协作又竞争的态势。在此情况下,科研评价方法需要与时俱进,传统的对被评价科研工作者、科

收稿日期:2013-09-03
资助项目:国家自然科学基金项目(71373019);中国博士后科学基金项目(2012M510520)
第一作者简介:王有国(1967—),男,湖北人,北京理工大学管理与经济学院博士研究生,研究方向:科研管理。

研机构和学术期刊静态的关注方式，已不能适应科技快速发展的需求，从学术关系的角度，结合现代信息可视化技术，从整体上动态地关注被评价对象及其研究内容，可以更好地发现其研究进展以及相互之间的推动、促进、依存、演化关系，从而有利于科技评价工作的进行。为此，国内外一些学者将学术关系和可视化技术应用到科研评价的研究实践中。如Boyack K W应用信息可视化技术对政府资助与科研论文产出关系进行评价[2]。Anastasios T应用科研合著关系可视化进行科研机构评价，并设计开发相应的评价系统辅助决策支撑[3]。Robert G应用引文可视化对文献进行筛选[4]，并结合聚类技术对领域专家进行评价。侯海燕应用信息可视化技术对科研工作者进行评价，识别科学学主流学术群体及其代表人物[5]。武汉大学科学评价研究中心邱均平设计开发的“科学评价管理信息系统”，采用科研合著关系、共词关系、共被引关系等学术关系辅助开展科研评价，同时提供与第三方可视化软件工具的接口，对构建的学术关系进行可视化展示[6]。中国科学技术信息研究所每年出版的《中国科技期刊引证报告》与《中国高被引指数分析》绘制了大量的科技期刊、科研工作者和科研机构共被引网络图，增强了读者对评价结果的理解[7-8]。

就目前可视化技术在科研评价中的应用来说，仍存在以下不足：与科研评价密切相关的可视化研究总体文献较少，缺少从方法层面开展的研究。应用在科研评价上的可视化方法以科研合著关系和共被引关系为主。在科研合著可视化上，更多强调的是合著者之间的整体关系、合作规律，而忽略了个体在群体中的贡献力，缺乏对合著模式、合著研究者个体研发能力微观层面的考察。在共被引可视化上，主要基于引文信息考察科研作者、机构、地区、期刊之间研究内容的相关性，缺乏对科研评价中“数量”、“质量”和“影响力”指标的反应。应用共被引可视化进行科研评价的有效性需要进一步提升。因此，研究科研评价的可视化方法，既有理论探索的意义，又有满足现实需求的价值。

2 研究方法与过程

本文研究科研评价的可视化方法，以数量指标、质量指标和相关性指标来综合反应评价结果。首先，对传统科研合著关系可视化方法进行改进，将科研合著者的论文署名顺序融合到可视化结果中，并结合文本挖掘，揭示其研究内容侧重，以此来评价个体研究者在群体中的贡献力。其次，基于文本挖掘技术对学术研究主体研究内容的关联性进行可视化分析，考察同一时间范围内科研工作者、机构、地区、期刊之间的相互关联性，评价其影响力，克服基于引文进行关联分析在揭示文献内容方面的不足。最后，构建科研工作者、机构、地区、期刊之间的年代引证关系，发现其特定时间阶段内研究内容或成果与以前的研究内容或成果的相关性，以及该阶段研究内容或成果与以后的研究内容或成果的相关性，进而评价其影响力。

2.1 学术研究主体识别

2.1.1 数据清洗与规范化

对文献集合中的科研工作者、科研机构、地区、期刊等学术研究主体进行识别抽取，利用人名词典、机构词典、地名词典进行规范化处理，合并相同主体，如相同作者、机构、地区的不同写法，不同作者、机构、地区的相同写法等。进而，建立规范后的研究主体与文献的隶属关系矩阵、学术研究主体与关键词同现矩阵，并按“学术研究主体×年代”对文献进行归类。

2.1.2 建立学术研究主体与文献隶属关系矩阵

记录每个学术研究主体在每个文献中出现的次序。如果同一主体在文献中出现多次，以第一次出现的次序为准。假设文献集合中规范后有 n 个研究主体，m 篇文献，构建研究主体与文献的隶属关系矩阵 $\boldsymbol{A}$ 如下。

$$\boldsymbol{A}=\begin{pmatrix} (b_1\ b_2\ b_3)_{11} & (b_1\ b_2\ b_3)_{12} & \cdots & (b_1\ b_2\ b_3)_{1j} & \cdots & (b_1\ b_2\ b_3)_{1m} \\ (b_1\ b_2\ b_3)_{21} & (b_1\ b_2\ b_3)_{22} & \cdots & (b_1\ b_2\ b_3)_{2j} & \cdots & (b_1\ b_2\ b_3)_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ (b_1\ b_2\ b_3)_{i1} & (b_1\ b_2\ b_3)_{i2} & \cdots & (b_1\ b_2\ b_3)_{ij} & \cdots & (b_1\ b_2\ b_3)_{3m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ (b_1\ b_2\ b_3)_{n1} & (b_1\ b_2\ b_3)_{n2} & \cdots & (b_1\ b_2\ b_3)_{nj} & \cdots & (b_1\ b_2\ b_3)_{nm} \end{pmatrix}$$

式中：$b_i=1$ 或 $b_i=0$，$\sum_{i=0}^{3}b_i=1$ 或 $\sum_{i=0}^{3}b_i=0$；$(b_1,b_2,b_3)_{ij}$

$=(1,0,0)_{ij}$ 表示主体 i 是文献 j 的第一著者；$(b_1,b_2,b_3)_{ij}=(0,1,0)_{ij}$ 表示主体 i 是文献 j 的第二著者；$(b_1,b_2,b_3)_{ij}=(0,0,1)_{ij}$ 表示主体 i 是文献 j 的第三或第三以后著者；$(b_1,b_2,b_3)_{ij}=(0,0,0)_{ij}$ 表示学术主体 i 不是文献 j 的著者。

2.1.3 建立学术研究主体与关键词同现矩阵

本文构建的共现矩阵 $\boldsymbol{B}$ 如下。

$$\boldsymbol{B}=\begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1k} & \cdots & e_{1l} \\ e_{21} & e_{22} & \cdots & e_{2k} & \cdots & e_{2l} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ e_{i1} & e_{i2} & \cdots & e_{ik} & \cdots & e_{il} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nk} & \cdots & e_{nl} \end{pmatrix}$$

式中：e_{ik} 为关键词 $Keyword_k$ 在主体 A_i 发表的文献中出现的频数；n 是学术研究主体总数；l 是文献组内所有关键词总数。

2.1.4 按‘学术研究主体×年代’对文献进行归类

把同一时间周期内发表的、具有相同研究主体的文献归为一类，每个类别表示为‘研究主体×年代’。

2.2 学术关系构建

2.2.1 构建科研合著关系

计算每个学术主体分别作为第一、第二、第三及以后合著者的文献数量之和，构建数量矩阵 $\boldsymbol{C}$ 如下。

$$\boldsymbol{C}=\begin{pmatrix} \sum_{j=1}^{m}(b_1)_{1j} & \sum_{j=1}^{m}(b_2)_{1j} & \sum_{j=1}^{m}(b_3)_{1j} \\ \sum_{j=1}^{m}(b_1)_{2j} & \sum_{j=1}^{m}(b_2)_{2j} & \sum_{j=1}^{m}(b_3)_{2j} \\ \vdots & \vdots & \vdots \\ \sum_{j=1}^{m}(b_1)_{ij} & \sum_{j=1}^{m}(b_2)_{ij} & \sum_{j=1}^{m}(b_3)_{ij} \\ \vdots & \vdots & \vdots \\ \sum_{j=1}^{m}(b_1)_{nj} & \sum_{j=1}^{m}(b_2)_{nj} & \sum_{j=1}^{m}(b_3)_{nj} \end{pmatrix}$$

式中：$\sum_{j=1}^{m}(b_1)_{ij}$, $\sum_{j=1}^{m}(b_2)_{ij}$, $\sum_{j=1}^{m}(b_3)_{ij}$ 分别表示学术主体 i 作为第一、第二、第三及以后著者的文献数量之和。

以学术主体与文献的隶属关系矩阵 $\boldsymbol{A}$，构建学术主体的合著关系矩阵 $\boldsymbol{AA}'$。其中，主体与文献的隶属关系矩阵 $\boldsymbol{A}$ 为 $n\times 3m$ 阶矩阵，$\boldsymbol{A}'$ 为 $3m\times n$ 阶矩阵，$\boldsymbol{AA}'$ 为 $n\times n$ 阶矩阵。假设 $\boldsymbol{AA}'=(a_{ij})_{nn}$，对于每个元素 a_{ij}，有：

$$a_{ij}=\boldsymbol{A}_i\boldsymbol{A}'_j=\left((b_1\ \ b_2\ \ b_3)_{i1}(b_1\ \ b_2\ \ b_3)_{i2}\cdots(b_1\ \ b_2\ \ b_3)_{im}\right)\times \begin{pmatrix} (b_1\ \ b_2\ \ b_3)'_{j1} \\ (b_1\ \ b_2\ \ b_3)'_{j2} \\ \vdots \\ (b_1\ \ b_2\ \ b_3)'_{jm} \end{pmatrix} \tag{1}$$

2.2.2 构建研究主体关联关系

利用研究主体与关键词同现矩阵，采用Tfidf特征表示方法[9]，对每个研究主体的文本特征进行表示，以夹角余弦作为研究主体之间关联度：

$$Sim(H_1,H_2)=\cos(\theta)=\frac{\sum_{k=1}^{n}w_{1k}w_{2k}}{\sqrt{\sum_{k=1}^{n}w_{1k}^2}\times\sqrt{\sum_{k=1}^{n}w_{2k}^2}} \tag{2}$$

式中：$H_1=(w_{11},w_{12},\ldots w_{1n})$；$H_2=(w_{21},w_{22},\ldots w_{2n})$；$w_i$ 为词 t_i 在研究主体 H 中出现频率的Tfidf函数值。

2.2.3 构建研究主体年代引证关系

按照‘学术研究主体×年代’对文献进行归类后，计算每一个类别下文献被其他类别下文献引用的数量之和，作为这一类的被引用数量，构建类别之间的引证关系矩阵 $\boldsymbol{D}$ 如下。

$$\boldsymbol{D}=\begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1j} & \cdots & e_{1k} \\ e_{21} & e_{22} & \cdots & e_{2j} & \cdots & e_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ e_{j1} & e_{j2} & \cdots & e_{jj} & \cdots & e_{jk} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ e_{k1} & e_{k2} & \cdots & e_{kj} & \cdots & e_{kk} \end{pmatrix}$$

式中：k 为归类后的类别数量；e_{ij} 表示类别 i 下所有文献引用了类别 j 所有文献的引用数量之和。需要指出的是，这种计算方法会出现两个类别相互引用数量都不为0的情况，且时间周期跨度越长，出现这种情况的几率越大。本文选择一年为周期，仍然会存在这种情况。在基于该矩阵进行可视化时，当出现两个节点互为引证的情况，仅保留引证数量较大的关系，忽略引证数量较小的关系，且不在可视化结果中显示。

2.3 学术关系可视化

2.3.1 可视化空间含义映射

本文应用二维空间中的网络图进行学术关系的

可视化含义映射，如表1。

2.3.2 可视化布局算法

应用Fruchterman T J和Reingold E M提出的复杂网络Fruchterman-Reingold算法进行可视化空间中网络节点布局[10]。该算法建立在粒子物理理论基础上，将无向图中的节点模拟成原子，通过模拟原子间的力场来计算节点间的位置关系。

同时，为了突出网络图中显著的学术关系，需要对网络图进行压缩，去掉关系不显著的连接线，识别关键信息。在进行基于文本的关联可视化上应用Pathfinder算法[11]，建立网络图中所有节点间最有效的连接路径。该算法对一个复杂网络中衡量数据相似性的关系进行简化，在所有可能的两点路径中只保留最强的连接，从而建立数据间最有效的连接路径。美国德雷克塞尔大学陈超美首先使用Pathfinder算法实现超文本链接网络聚类[12]，并在其设计开发的可视化工具CiteSpace中进行固化[13]。

3 实证应用

3D打印技术被认为是近二十年来制造领域的重大突破，得到国内外众多科研工作者和新闻媒体的广泛关注。为此，本文采用文献[14]的检索词，在SCI论文数据库中，检索1998—2012年间，题目包括相关检索词的论文数据，共712篇。为避免因数据整理不完全而引起的评价结果不准确，选择主要国家(地区)作为评价对象，从科研合著关系、关联关系、年代引证关系，揭示其研究成果在同领域的数量、质量和影响力情况。

图1为论文涉及的53个国家(地区)间的合著关系，图中标记了每个国家署名顺序分别为第一、第二、第三及以后的论文数量。总体来看，美国、中国、德国、英国和韩国论文数量位居前五；考虑到论文署名顺序，韩国作为第一作者的论文数量超过了英国，英国的第二作者论文和第三及以后作者论文数量占总体论文数量比较高，且与该国合著关系较强的国家为美国、中国和德国；美国的第二、第三及以后作者的论文数量均为11，但与124篇第一作者论文数量相比还是较少的。其他国家(地区)的论文数量与这5个国家相比要少得多。

进一步，将发文数量较多的前20位国家(地区)挑选出来，绘制其文本关联可视化图形，并标注每个国家(地区)涉及最多的五个学科类别和国家(地区)间的文本关联强度数值，如图2。把那些不具有合著关系，但研究内容仍然存在明显关联性的国家(地区)挑选出来，如图2圈定的几个群体，每个群体内部国家(地区)间研究内容相互影响，具有一定的竞争关系。

再从图3的历史引证关系看，美国、德国、英国、加拿大在2008年的论文被引用数量较多，对之后研究的影响相对显著，这与数据检索的时间范围有一定关系；同时，也应注意到，各个国家之间的自引要明显多于他引，比如我国2008年、2009年、2011年、2012年的论文引用，美国2008年、2009年、2010年和美国2011年、2012年的论文引用等。以此类推，可对每个国家(地区)各年发表论文与之前或之后论文的引证关系和相互影响进行揭示与评价。

4 结 语

作为现有科研评价方法的辅助手段，论文构建

表1 学术关系可视化含义映射

学术关系	可视化空间含义映射						评价内容
	节点大小	节点颜色	节点文字	连线粗细	连线反向	连线文字	
科研合著关系	与文献数量多少成正比	红、绿、黄色分别表示署名第一、第二、第三及以后的文献数量	研究主体名称及其研究重点关键词、学科、技术类别	与合著数量成正比	无意义	合著文献数量	数量指标
研究主体关联关系	同上	同上或无意义	同上	与关联强度成正比	无意义	关联强度数值	横向影响力
研究主体年代引证关系	同上	同上或无意义	研究主体名称	与被引用数量成正比	与引用方向相同	被引用数量	质量指标、纵向影响力

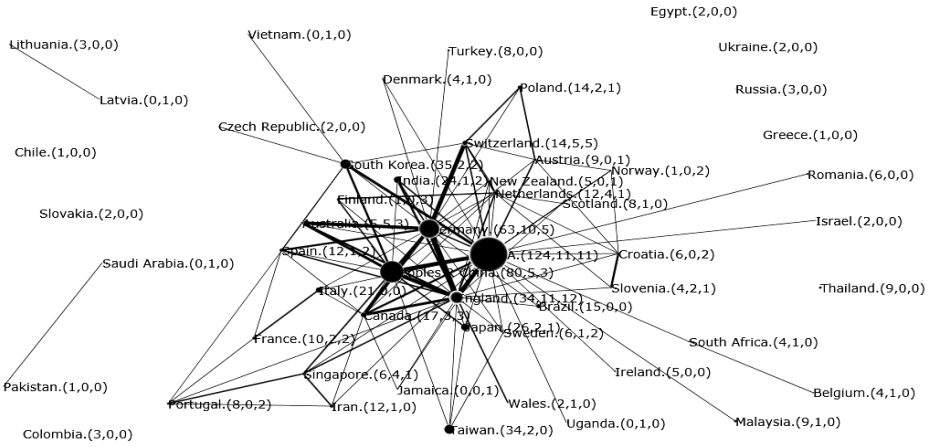

图1 3D打印技术SCI论文国家合著关系

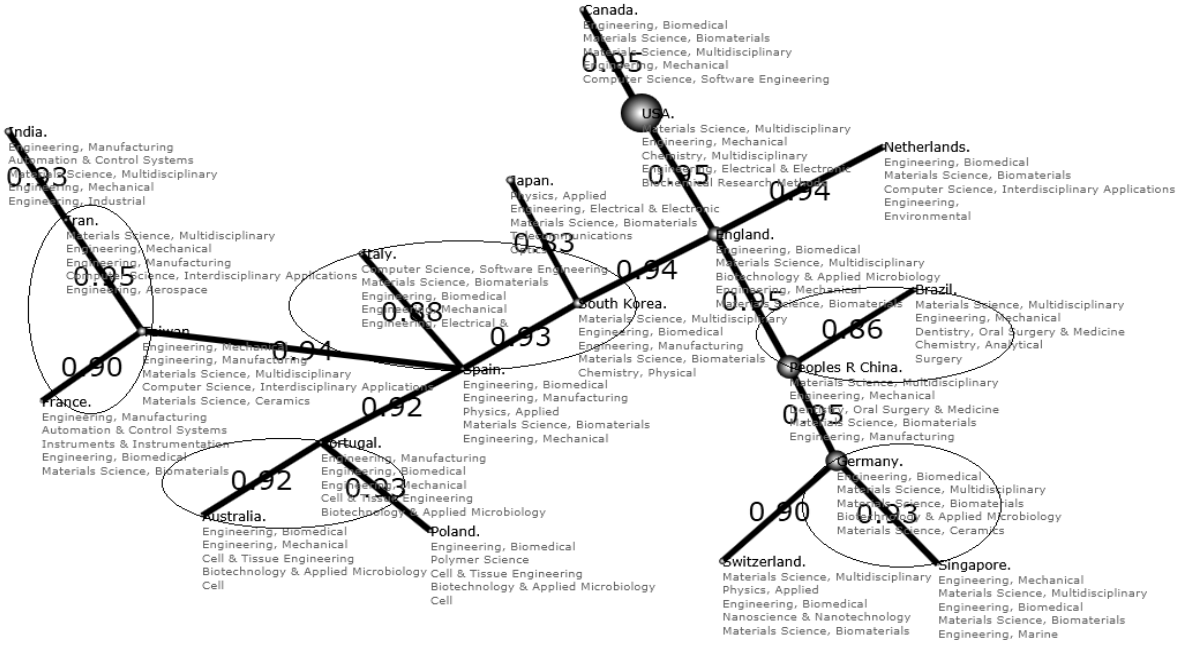

图2 3D打印技术SCI论文国家关联关系

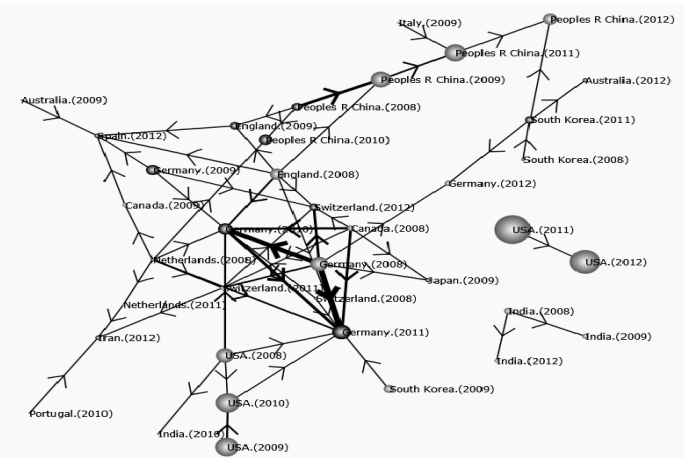

图3 3D打印技术SCI论文国家年代引证关系(设定引证数量大于1)

科研合著关系、研究主体关联关系、研究主体年代引证关系，对构建结果采用网络图进行可视化呈现，以此来揭示特定研究领域被评价对象科研成果的数量、质量和影响力。评价的有效性与数据质量有着直接的关系，如同一作者、机构、地区的不同写法，不同作者、机构、地区的相同写法，都会对评价结果产生影响。在应用实践中，与其他基于文献计量学的评价方法一致，需要进行数据的规范化处理，并综合其他评价方法，使评价结果尽可能反映被评价对象的真实情况。在今后的研究中，如何改进关联度的计算方式，更加清晰地表现学术关联，与大数据应用相结合，扩大实证研究的数据规模，以及基于本文评价方法进行相关软件工具的设计开发，都是值得进一步研究的问题。

参考文献

[1] 陈敬全. 科研评价方法与实证研究[D]. 武汉:武汉大学,2004.

[2] Boyack K W, Borner K. Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers[J]. Journal of the American Society for Information Science and Technology, 2003,54(5):447-461.

[3] Anastasios T, Sgouropoulou C, Xydas I. Academic research policy-making and evaluation using graph visualisation[C]. Kastonia: Proceedings of the 15th Panhellenic Conference on Informatics, 2011.

[4] Robert G, Cody D, Ben S. Evaluating visual and statistical exploration of scientific literature networks[C]. Pittsburgh: IEEE Symposium on Visual Languages and Human-Centric Computing, 2011.

[5] 侯海燕,刘则渊,陈悦,等. 当代国际科学学主流学术群体及其代表人物[J]. 科学学研究,2006,24(2):161-164.

[6] 邱均平,杜晖. 科学评价管理信息系统构建[J]. 图书情报知识,2013(1):56-62.

[7] 中国科学技术信息研究所. 中国科技期刊引证报告[R]. 北京:北京科学技术出版社,2013.

[8] 中国科学技术信息研究所. 中国高被引指数分析(2012年版)[R]. 北京:北京科学技术出版社,2013.

[9] 史忠植. 知识发现[M]. 北京:清华大学出版社,2002.

[10] Fruchterman T J, Reingold E M. Graph drawing by force directed placement[J]. Software Practice and Experience, 1991,21(11):1129-1164.

[11] Schvaneveldt R W, Dearholt D W, Durso F T. Graph theoretic foundations of pathfinder networks[J]. Computers and Mathematics with Applications, 1998,15(4):337-345.

[12] Chen C. Generalised similarity analysis and pathfinder network scaling[J]. Interacting with Computers, 1998,10(2):107-128.

[13] Chen C. CiteSpace: Detecting and visualizing emerging trends and transient patterns in scientific literature [J]. Journal of the American Society for Information Science and Technology, 2006,57(3):359-377.

[14] 刘红光,杨倩,刘桂锋,等. 国内外3D打印快速成型技术的专利情报分析[J]. 情报杂志,2013,32(6):40-46.

(责任编辑 孙 兰)

Visual Method of Academic Relationship for the Research Evaluation

WANG Youguo[1], LIU Yuqin[2], WANG Xuefeng[1]

(1. School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China;2. Academic of Printing and Packaging Industrial Technology, Beijing Institute of Graphic Communication, Beijing 102600, China)

Abstract: For the purpose of overcoming the current defects about visualization technology application of research evaluation, the paper improves the traditional co-author visualization, integrates association visualization, and constructs citation visualization to convey the quantity, quality and impact of the object being evaluated from the perspective of academic relations. The evaluation method proposed in the paper can be used for researchers, institutions, regions and journals which engaged in some specific field. The research background and process are elaborated. At last an empirical illustration about 3D print is put forward.

Key words: research evaluation; academic relationship; visualization