情报学报 ISSN 1000-0135
第27卷第1期69-74，2008年2月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol.27 No.1，69-74 February 2008

关于共被引分析方法的再认识和再思考[1)]

邱均平 马瑞敏 李晔君
（武汉大学信息管理学院，武汉 430072）

摘要 共被引（Co-citation）分析方法自1973年提出后便产生了很广泛影响，大家一直以来都按约定俗成的方式去处理共被引矩阵，但是最近对该方法的讨论非常激烈。核心问题围绕“矩阵转化的方法”而展开。在认识他们争论的同时，我们对此也提出了自己的想法和思考。我们从“共被引矩阵作为临近矩阵”这一角度，认为对角线应该采用该作者与其他作者的最大共被引次数+1来表示；同时我们从稳定性和程序的统一性等方面探讨了欧几里得距离的平方这一标准化算法的优越性。最后我们用实例来证实了自己的想法的可行性和准确性。以期对以后该方法的使用提供参考。

关键词 共被引 皮尔逊相关系数 聚类分析 多维尺度分析 可视化 Pajek

Reconsiderations on Co-citation Analysis

Qiu Junping Ma Ruimin and Li Yejun
（*School of Information Management*, *Wuhan University*, *Wuhan*, 430072）

Abstract Co-citation has been affecting wildly since put forward in 1973, most of us deal with the co-citation matrix in the ways established by usage, but recently the discussions on it are hot. The core problem is how to convert the matrix. At the same time, we also put forward our thoughts and ideas. We think the diagonal value should be put with maximum + 1 in terms of cocitation matrix as a proximity one. Meanwhile, we discuss the advantages of squared Euclidean distance from the aspects of stability and oneness of the processing program. In the end, we take an example to approve our thoughts and hope these can supply some beneficial reference to further use of this method.

Keywords co-citation, Pearson's correlation coefficient, cluster, MDS, visualization, Pajek

1 引 言

共被引分析是由Small和Marshakova于1973年同时分别提出的。从此对共被引分析的研究和实践在科学计量学领域内广泛展开。1981年，White和Griffith将其扩展到作者共被引分析（ACA），它对于探讨学科结构有着积极的开创意义。比如1989年，White和McCain通过ACA认为图书馆和信息科学（Library and Information Science，LIS）学者可以主要分为两种：科学计量学（文献计量学）学者群和信息检索学者群，这就为LIS学科的基本定位和以后主要发展方向提供了数据咨询。现在，共被引分析在各个学科领域都有着广泛的应用。但几乎国内外所有学者都是按照统一的方法模式来进行共被引分析，即第一步构造共被引矩阵；第二步将该矩阵转化为相似系数矩阵，方法大多为皮尔逊相关系数法（Pearson's Correlation Coefficient）；第三步是进行聚类

收稿日期：2006年11月10日

作者简介：**邱均平**，男，1947年生，武汉大学信息管理学院教授、博士生导师，主要研究方向：信息管理与科学评价，知识管理与竞争情报。**马瑞敏**，男，1983年生，武汉大学信息管理学院2006级博士，主要研究方向：信息计量与科学评价。E-mail：ruimin.ma@yahoo.com.cn。**李晔君**，女，1984年生，武汉大学信息管理学院2006级硕士，主要研究方向：信息计量与科学评价。

1） 国家社科基金重点项目“我国人文社科研究评价体系的构建与实证分析”（05AZX004）的成果之一。

(Cluster)和多维尺度分析(MDS)。这基本是大家默认的方法,在我国许多相关教材和相关论文中也多是采用这样的方法。

但是我们关注到,从2003年开始,关于共被引分析方法的再次讨论又在科学计量学学者间展开了。这次讨论源自于Ahgren、Jarneving和Rousseau共同发表的《共被引相似性测度的必要条件——特别以皮尔逊相关系数为证》一文,他们在论文中提出共被引相似性测度的两个必要条件①,从而提出对皮尔逊相关系数的质疑,建议用Cosine和Chi-Squared Distance两种方法代替皮尔逊相关系数②。此后就皮尔逊相关系数引发的问题便一直成为科学计量学家关注的热点。科学计量学界最高奖普赖斯奖获得者——White, Rousseau, Leydesdorff等权威专家都加入了此次讨论。

在2005年前,这场争论主要集中在共被引矩阵转化为相似系数矩阵方面,特别是对皮尔逊相关系数是否适合应用到共被引分析这一问题展开,由此也引发共被引矩阵对角线如何取值的讨论。这段时间的争论主要是在支持保持皮尔逊相关系数测度方法的Bensman和White同反对继续使用皮尔逊相关系数测度方法而采用其他方法的Ahgren等学者之间展开。我们可以发现他们虽然争论,但有一个共同点:都认同原始共被引矩阵应该转化为相似系数矩阵,只不过是使用不同方法而已。而在2006年,Leydesdorff发表了《共现分析及在信息科学中的应用》一文,在该文中他提出原始共被引矩阵根本不应该进行转化。虽然现在还没有看到更新的相关文献的对此评论,但是Leydesdorff这篇文章无疑将带来对共被引分析的更大范围的讨论。

我们对此也进行了一些思考,尤其对讨论的突出焦点问题进行了思考,主要从"对角线取值","Pearson's还是其他"这两个问题展开,并用实例来证明我们提出的观点。

2 相关知识简介

由于共被引分析用到许多科学计量学和统计学基本知识,包括矩阵、相似性测度等,为了更好地理解后面的讨论,在这里将它们做简要的介绍。

(1)临近矩阵(Proximity Matrix)。

定义Ⅰ 其定义为:一组表示两两目标之间相似程度或者不相似程度的数字组成的矩阵。它是沿对角线对称的方矩阵。

多维尺度分析中输入矩阵必须是临近矩阵,聚类分析也要将矩阵通过不同的测度方法转化为临近矩阵。

(2)共被引矩阵(Co-citation Matrix)

定义Ⅱ 其定义为:两篇文献被别的文献同时引用,并以引用它们的文献数量作为共引强度。

其最早由Small提出的,它是完全对称的矩阵,也是临近矩阵,其对角线选择默认值。具体形式如表1所示。其中A, B, C, D表示论文、著者、期刊和学科等研究对象,$NO.(ij)$表示i和j两研究对象之间的共被引强度,$NO.(ij) = NO.(ji)$,对角线(即$i = j$)为默认值,在SPSS中用"."来表示。

表1 对称共被引矩阵

	A	B	C	D
A		NO.(AB)	NO.(AC)	NO.(AD)
B	NO.(BA)		NO.(BC)	NO.(BD)
C	NO.(CA)	NO.(CB)		NO.(CD)
D	NO.(DA)	NO.(DB)	NO.(DC)	

共被引矩阵是表示的是两目标之间的相似程度的矩阵,即两者数字越大表明两者关系越近,越小表面两者关系越远。

(3)皮尔逊相关系数(Pearson's Correlation Coefficient)。其表达式为:

$$r_{XY} = \frac{S_{XY}}{S_X \times S_Y}, \quad (1)$$

其中,S_{XY}表示两变量的协方差,S_X表示变量X的标准差,S_Y表示变量Y的标准差。

其值可正可负,表示两变量之间的相似程度。在传统共被引矩阵转化方面该方法为默认方法。

(4)Cosine测度。其表达式为:

① 具体内容如下:(1)对于相似性测度,变量A和变量B的相关系数$s(A,B)$在加入0-模块后不能减小;(2)未加入0-模块前,如果$s(A, B) > s(C, D)$,那么加入0-模块后,这种关系也仍要保持。从公式Ⅰ来看,皮尔逊相关系数在一些情况下是不符合这两个条件。

② 标准化后的共被引矩阵的数据类型看成是连续变量(Interval),而Chi-Squared Distance是用于计数变量(Counts),所以后文对Chi-Squared Distance讨论没有展开。主要讨论的是Cosine测度,它可用于对连续变量的测度。

$$S(X,Y)=\frac{\sum_i}{\sqrt{\sum X_i^2\sum Y_i^2}},i=1,2,L,s \quad (2)$$

其值为正,也表示两变量的相似程度。在此次争论中,该测度被建议取代皮尔逊相关系数测度。

(5)欧几里德平方距离。其表达式为:

$$D^2(X,Y)=\sum(X_i-Y_i)^2,i=1,2,L,s \quad (3)$$

其值为正,表示两变量的不相似程度。

3 对角线

关于对角线取值问题一直也是争论不休,主要是两个方面:一是SPSS的一些程序对默认值无法处理,必须要填入相应的数字,比如聚类分析(Cluster)和多维尺度分析(MDS);二是默认值的取值对于计算皮尔逊相关系数等方面有影响,不同的取值会导致不同的相关系数。这两个方面事实是密切联系的。按照传统方法我们都会把原始矩阵转化为Pearson相关矩阵,这样对角线就自动变为1,也解决了程序的限制问题。从表面看来,在处理数据方面非常方便,但往往忽视了其统计学和科学计量学的意义。

White和Griffith最初是将对角线值定为:排序前三的共被引频次之和/2;McCain将其定为:默认值,这也是影响最广泛的一种对角线确定方法;Ahlgren等认为应该使用自己与自己实际共被引次数;White则建议使用最大值来确定对角线的值。

我们认为根据共被引原理来看,Ahlgren等的建议是合理的。但是从临近矩阵的定义(定义Ⅰ)来看,White的建议则更正确,共被引矩阵本来就是考察各对象之间亲疏关系的临近矩阵,只不过研究的目标选定了有特殊意义的作者、论文、期刊、学科等而已。从这点出发,我们自然而然的认为自己和自己的关系最亲近,所以应该是该作者与其他作者共被引频次中最高的。表达方式可以是最大值。但我们认为为了突出自己与自己的亲密关系,可以用最大值+1来凸显。所以我们大体倾向于White的提法,但可做适当的调整。

4 Pearson's还是其他

4.1 重申共被引矩阵标准化的重要性

由于共被引强度受学科、专业甚至研究方向的影响很严重,所以组成的矩阵数据差别大,比如,科学计量学者之间共被引强度高而与信息检索学者共被引强度低。这样相当于变量单位不同而造成了数据相差悬殊的现象。而标准化则可以缩减这样的差距,减少突出数据的影响,在以后矩阵的运算中能更好的表现出变量间的关系。至于如何标准化,在4.3节将详细说明。

4.2 输入矩阵形式

由于Leydesdorff选取的数据样例为美国各大城市之间的距离,他认为用Pearson's转化后的矩阵进行MDS分析时与实际情况不符合。但我们认为这并不是否定共被引矩阵转化的充分理由。因为地理数据和共被引数据是有差别的。对于地理数据,A城市离B城市近,B城市离C城市近,那么A城市和C城市也离的近,这种关系推理很明显,是必然的。但对于共被引数据这种关系并不是必然的,有时甚至差别很大。比如A作者和B作者共被引高,B作者和C作者共被引也高,但是不一定推出A作者和C作者共被引也高。所以这就需要用一种测度方法将矩阵转化,进一步发现各变量之间的关系(在数据标准化基础上)。所以我们认为共被引矩阵应该转化,但是可以把原始矩阵输入在SPSS中自动实现转化,无需把转化后的矩阵作为输入矩阵。至于到底如何实现这一目的我们将在下文详细展开。

4.3 SPSS相关程序说明

首先强调的是,我们使用的是SPSS 11.5版本。

对于聚类分析,我们最主要的是使用了图1的对话框,我们可以看到,在Cluster Method中可选择不同的聚类方法,包括最远距离法,Ward's法等。在Measure中我们可以清楚看出各种测度方法,比如Squared Euclidean distance, Cosine, Peason correlation等。这些测度方法将原始共被引矩阵转化为目标矩阵(比如Peason相关系数矩阵等),然后根据给定的Cluster Method来进行聚类并输出结果。另外,对于4.1节提出的标准化方法,可以在图1对话框中一并实现,即在Standardize下选择某种方法,比如z分数,0~1数值范畴等。

对于多维尺度分析(Multidimentional Scaling (PROXSCAL)),首先强调的是经标准化的新矩阵不是临近矩阵,需要转化。但是也没有必要将标准化后的矩阵作为输入矩阵。我们可以借助SPSS自动实现。具体操作如下:首先要选择数据形式为"从数据中创建临近矩阵(Create proximities from data)"。然

后在图2所示的对话框中，我们可以看到各种转化方法。另外标准化方法也可以同时实现。需要注意的是，在转化的方法中没有 Pearson's 相关系数，只有 Euclidean distance 和 Squared Euclidean distance 这样的非相似性测度方法。这样在输入原始矩阵后也可自动实现矩阵标准化和非相似性测度(Squared Euclidean distance)转化。这和先输入标准化矩阵，然后转化为非相似性测度矩阵是一样的效果。

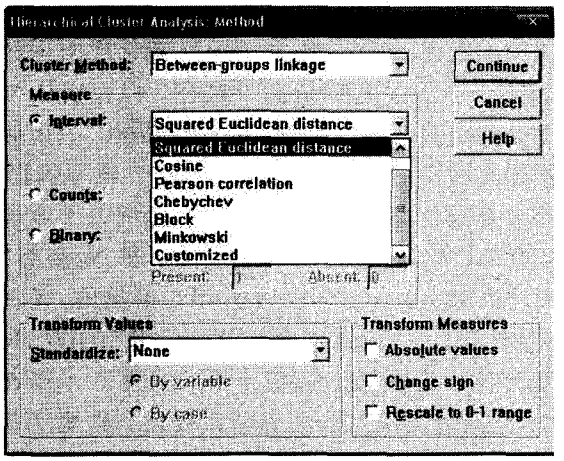

图1 聚类分析的方法对话框

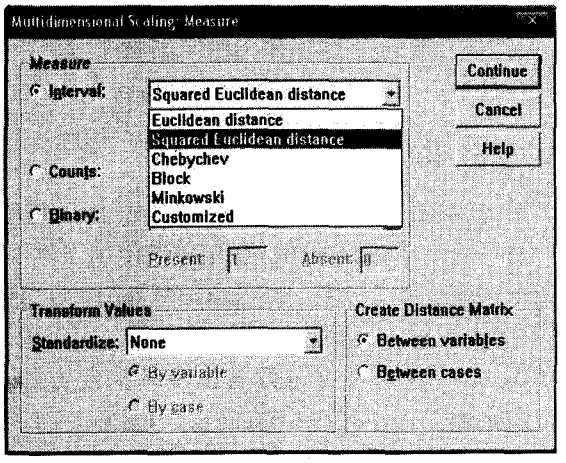

图2 多维尺度分析的测度对话框

4.4 用 Squared Euclidean distance 代替 Pearson's

Pearson's 只是一种测度变量相似性的方法，是为了更好地发现变量之间的关系，而许多方法都可以取代它。我们认为 Squared Euclidean distance 是代替它的最好方法。理由如下：

(1)Pearson's 相关系数矩阵自身的确存在问题。其不适合对有0-模块的矩阵进行转化，虽然 White 等一再强调，共被引矩阵不应该存在0-模块，并且给出了许多实例。但是我们认为这还是不具有普遍性，难免遇到0-模块，尤其是在我国的研究者之间。在这里，我们再次肯定 Ahlgren 等提出的相似性测度的两个必要条件是非常正确的。虽然是针对相似性测度提出，但是其原理对于非相似测度同样正确。我们用 Squared Euclidean distance 测度方法是满足两个必要条件。公式Ⅲ可以看出即使在矩阵后加入0-模块，$D^2(X, Y)$ 的值保持不变，说明 Squared Euclidean distance 测度方法和 Cosine 测度(见公式Ⅱ)方法都具有很好的稳定性，只不过表示的变量之间的关系不同而已，一个为非相似性测度，一个为相似性测度。

(2)程序的统一性。不论聚类还是多维尺度分析，我们都可以将原始矩阵直接作为输入矩阵，事先不需要任何转化。对于聚类，我们可以根据图1所示将原始矩阵按照 Squared Euclidean distance 测度方法和给定的标准化方法转化为 Squared Euclidean distance 非相似矩阵。对于多维尺度分析，我们也可按照同样的方法将原始矩阵转化为 Squared Euclidean distance 非相似性矩阵。这样我们不管是聚类还是多维尺度分析，实际操作的矩阵都是 Squared Euclidean distance 非相似性矩阵。如果用 Cosine 替代的话，程序的统一性难以保障，因为对于多维尺度分析，必须输入 Cosine 相似矩阵，而这又需要对原始矩阵进行转化(包括标准化)，程序又显得繁琐。

(3)另外着重强调一点，我们从图1也注意到，对于聚类分析也完全没有必要把原始矩阵转化为 Peason's 相关系数输入矩阵，否则我们将选择什么样的测度来继续转化矩阵？有画蛇添足之意。

4.5 实例

为了检验该方法的可行性和正确性，我们选取了我国影响力较大的6位科学计量学学者和6位信息检索学者进行著者共被引分析(ACA)。利用的是 CNKI 的中国期刊全文数据库 1990 年 1 月 1 日至 2006 年 10 月 23 日的数据。我们将原始共被引矩阵直接输入进行聚类和多维尺度分析。仍然需要强调几点：一是对角线我们定为最大值 + 1；二是标准化方法我们选择 z 分数；三是聚类方法选择 between-groups linkage。四是在多维尺度分析中我们要将临近转化方式选定为 Ordinal。五是对于科学计量学学者和信息检索学者之间的共被引文献进行了详细分析，去除了对于 LIS 教育和综述方面的文章，进行了必要的数据调控。图3为原始共被引矩阵，图4为聚类分析结果，图5为多维尺度分析结果。

	王崇德	邱均平	罗式胜	蒋国华	赵红洲	梁立明	张琪玉	赖茂生	陈光祚	曾民族	侯汉清	焦玉英
王崇德	230	229	72	20	7	6	7	6	5	3	3	1
邱均平	229	230	112	17	9	24	1	6	10	7	1	2
罗式胜	72	112	113	15	2	10	2	9	2	2	1	0
蒋国华	20	17	15	31	26	30	0	1	0	0	0	0
赵红洲	7	9	2	26	27	3	1	0	1	0	0	0
梁立明	6	24	10	30	3	31	0	0	0	0	0	0
张琪玉	7	1	2	0	1	0	148	30	40	9	147	19
赖茂生	6	6	9	1	0	0	30	32	31	17	17	9
陈光祚	5	10	2	0	1	0	40	31	41	27	14	17
曾民族	3	7	2	0	0	0	9	17	27	28	3	5
侯汉清	3	1	1	0	0	0	147	17	14	3	148	7
焦玉英	1	2	0	0	0	0	19	9	17	5	7	20

图 3　原始共被引矩阵

从图 3 来看，该矩阵的对角线的值为某作者与其他作者共被引强度最大数 + 1，并且我们发现的确存在 0-模块。所以不适宜用 Pearson's。

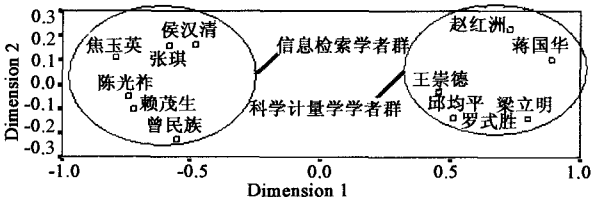

图 4　为聚类分析结果

我们直接将原始矩阵输入，并利用 Squared Euclidean distance 测度将其转化为非相似临近矩阵。图 4 的聚类结果是和实际相符的。将科学计量学学者群和信息检索学者群很清晰的划分开来。并且每个学者群内的结构也比较清晰，比如科学计量学学者群中，王崇德、邱均平和罗式胜都写过关于科学计量学方面的专著，被引也一直都很高，从图 3 也可看出，他们之间的共被引次数相对较高。他们被分在二级小类中是完全合理的。

同聚类分析一样，我们也直接将原始矩阵输入进行多维尺度分析，其可视化结果也是非常清晰的。我们可以很明显的看到两个不同的学者群分布在同一坐标的不同两侧，每个学者群内部学者的距离也

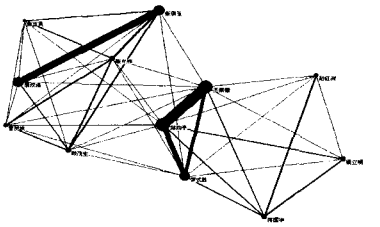

图 5　为多维尺度分析结果，Normalized Raw Stress = 0.0012

符合实际情况，比如陈光祚和赖茂生离的很近，王崇德和邱均平、罗式胜离的很近。

从以上实证，可以看出我们提出的方法能够较好地进行共被引分析，不论从聚类角度还是多维尺度分析的可视化结果来看，能很好的区分两个学者群，并且其与实际情况是相符的。

另外随着社会网络分析可视化方法的发展，一些用于数据可视化的软件在国外涌现，Pajek 便是其中非常优秀的工具之一，其用于学术研究目的其使用是免费的。多维尺度虽然可以较好的观察到变量间的关系，但是无法表现他们之间的强弱，而 Pajek 可以较好的弥补这一缺陷。图 6 是使用 Pajek 可视化的结果，其中线的粗细可看出学者之间的共被引强度，比如邱均平和王崇德之间的线最粗，说明他们的共被引强度最高。节点的大小表示的是对角线的值，表明作者与自己的共被引强度，实际暗含着该作者的影响力大小。

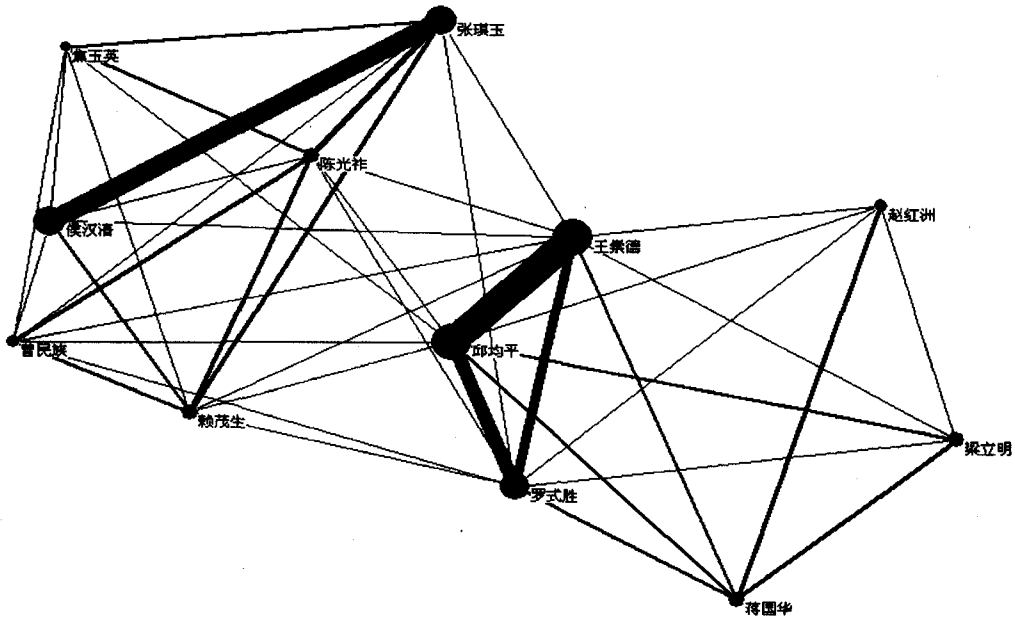

图6 使用Pajek可视化共被引矩阵

5 结论和讨论

我们认为对于同被引矩阵需要有新的认识，上文正是对对角线问题和矩阵如何转化问题进行了重点讨论。总结我们的观点，核心两点如下：

(1)对角线问题需要重新审视。对角线取值大小会影响相似性和非相似性测度的值，也必然会影响聚类分析和多维尺度分析的结果。我们提出用该作者与其他作者共被引频次的最大值+1来表示。

(2)结合Leydesdorff、White和Ahgren等学者的不同观点，我们深入分析了到底如何处理共被引矩阵问题，我们的解决方法是将其直接输入，然后同时将数据标准化和Squared Euclidean distance测度转化在SPSS的聚类分析和多维尺度分析中自动完成，程序简单统一而又保证了其可行性。

最后，希望本文能够对以后共被引分析的深入发展提供一些有益的参考。

参 考 文 献

[1] Ahlgren P, Jarneving B, Rousseau R. Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. Journal of the American Society for Information Scienceand Technology, 2003, 54(6): 550-560.

[2] Ahlgren P, Jarneving B, Rousseau R. Author cocitation and Pearson's r. Journal of the American Society for Information Science and Technology, 2004a, 55(9): 843.

[3] Ahlgren P, Jarneving B, Rousseau R. Rejoinder: In defense of formal methods. Journal of the American Society for Information Science and Technology, 2004b, 55(10): 936.

[4] Bensman S J. Pearson's r and Author Cocitation Analysis: A commentary on the controversy. Journal of the American Society for Information Science and Technology, 2004, 55(10): 935-936.

[5] Egghe L, Rousseau R. Introduction to Informetrics. Amsterdam: Elsevier, 1990.

[6] Leydesdorff L, Vaughan L. Co-occurrence Matrices and Their Applications in Information Science: Extending ACA to the Web Environment. Journal of the American Society for Information Science and Technology, 2006, 57(12): 1616-1628.

[7] White H D. Author cocitation analysis and Pearson's r. Journal of the American Society for Information Science and Technology, 2003, 54(13): 1250-1259.

[8] 黄润龙. SPSS软件实用教程. 北京：高等教育出版社，2004.

[9] 罗式胜. 文献计量学概论. 广州：中山大学出版社，1994.

[10] 张文彤. SPSS 11.0统计分析教程(高级篇). 北京：北京希望电子出版社，2002.

(责任编辑 王建平)