学术研究主体与研究内容间的关联关系可视化方法*

周 杰 刘玉琴 曾建勋
(中国科学技术信息研究所 北京 100038)

【摘要】提出一种学术研究主体与研究内容间的关联关系可视化方法,该方法通过改进对应分析的多元统计分析方法,进行学术研究主体之间、研究内容之间以及学术研究主体和研究内容之间的关联度计算,采用网络图进行关联关系的可视化表示。阐述该方法研究的背景、方法和过程。最后,以图书情报与数字图书馆学科中信息可视化、知识图谱相关的学术论文数据为例进行实证应用。

【关键词】对应分析 关联关系 可视化
【分类号】TP391

Visualization Method of Correlative Relationship Between Academic Research Body and Content

Zhou Jie Liu Yuqin Zeng Jianxun
(Institute of Scientific & Technical Information of China, Beijing 100038, China)

【Abstract】The paper proposes a visualization method of correlative relationship between academic research body and content. Correspondence analysis is improved to calculate correlation among research bodies and contents, then network graph is used to show the output, and the research background, method and process are also elaborated. At last, an empirical illustration about information visualization and knowledge map in the intelligence and digital library science is put forward.

【Keywords】Correspondence analysis Correlative relationship Visualization

1 引 言

从信息可视化的角度研究海量学术资源中潜在的学术关系,从整体上动态关注学术研究主体、研究内容的进展以及相互之间的推动、促进、依存、演化关系,不仅有利于国家科技发展战略的制定,学术研究机构的科技管理,学术研究者个人研究能力的提升,还有助于学术资源的整合与管理,丰富学术信息服务商、服务机构的服务内容与方式。

为丰富现有学术关系分析方法,本文构建学术研究主体与研究内容之间的关联关系可视化表示,将对应分析的多元统计分析方法应用于关联度的计算过程;然后,对学术研究主体、研究内容及其关联强弱进行可视化空间含义映射,采用网络图进行可视化呈现,以此来揭示学术研究主体、研究内容之间以及学术研究主体和研究内容之间的关联关系。

收稿日期:2012-10-18
收修改稿日期:2012-11-14
*本文系国家自然科学基金项目"基于海量数字资源的科研关系网络构建研究"(项目编号:71203208)和中国博士后科学基金项目"学术关系可视化方法与系统研究"(项目编号:2012M510520)的研究成果之一。

2　研究背景

当前,国内外众多学者采用各种方法,从不同角度对学术关系进行研究,如科研合作、文献共词、文献引证、学术关联等。其中,针对学术关联的研究可归纳为围绕作者、机构、地区等学术研究主体间的关联[1-3],围绕主题、学科、技术领域等学术研究内容间的关联[4-6],以及基于学术关联开展的技术预测、学术评价、数字图书馆信息增值服务等[1,7,8]。

在实际需求分析中,研究者不仅关注学术研究主体之间、研究内容之间的关联,而且也关注学术研究主体与研究内容之间的关联。比如,在进行学术研究机构的关联分析中,不但要知道各个机构研究内容的关联性,还要明晰这些机构各自的研究重点、技术优势。为此,一些研究人员将多元统计分析中的对应分析引入到学术关联分析中。Doré 等[9,10]利用该统计方法对 48 个国家的 18 个科研领域的期刊文献进行分析,用以发现各个国家在各领域的科研优势,并应用该方法进一步对专利文献进行分析挖掘。Bhattacharya 等[11]利用该方法研究 INSPEC 数据库中物理学领域 1990 – 1995 年间 20 个主要国家和研究主题间的关联。Anuradha 等[12,13]应用该方法对印度在各学科的国际合作中的特点进行分析,揭示印度国际合作中国家与技术领域的关联性,后将该方法与聚类分析结合进行科技文献的分析。Iribarren – Maestro 等[14]应用该方法对马德里大学的 10 个技术领域的科技文献引证和合著的关联性进行揭示。刘玉琴[15]结合文本挖掘与对应分析进行通信领域中国专利文献分析,用以挖掘领域内的研究机构与技术领域的关联性。

在关联结果的表示上,基于经典对应分析的关联关系采用二维坐标图进行样本和变量的关联性表示,这种表示存在两方面不足:

(1)由于对应分析方法本身将多维空间中点的布局压缩到二维空间,导致信息不完备;

(2)当样本和变量较多时,坐标点之间距离较小,坐标点相互重叠,图形复杂度增加,不易对分析结果进行阅读和理解。

在各种可视化技术中,基于复杂网络算法的网络图是操纵大型网络结构数据且在学术研究主体的合作、关联、引证等学术关系分析中应用广泛的技术之一。如国外的文献分析软件 VantagePoint[16] 和 Thomson Data Analyzer[17] 均采用网络图进行关联结果的可视化表示。

基于以上背景分析,本文研究将对应分析方法与复杂网络相结合,设计并实现一个能够全面反映学术研究主体之间、研究内容之间以及学术研究主体和研究内容之间关联关系的可视化方法。

3　研究方法与过程

对应分析将样本信息与变量信息统一起来,以二维坐标图中的点及其距离来分析样本之间、变量之间以及样本和变量之间的关联性。在学术研究主体与研究内容的关联可视化上,将学术研究主体作为样本信息,研究内容作为变量信息;采用对应分析法将其映射为多维空间中的点;以适当的变换将点的距离值转化为点的关联度;进而,结合网络图进行关联结果的可视化表示。以下为该方法的具体研究过程。

3.1　关联度计算

在关联度计算上,基于对应分析方法进行算法改进,首先构建学术研究主体和研究内容(技术类别或技术关键词)之间的同现频数矩阵 $X=(x)_{n\times m}$,然后进行矩阵变换[18],公式如下:

$$z_{ij}=\frac{x_{ij}-x_{i.}\,x_{.j}/\sum_{i=1}^{n}\sum_{j=1}^{m}x_{ij}}{\sqrt{x_{i.}\,x_{.j}}}(i=1,2\cdots n;j=1,2\cdots m)$$

得到矩阵 $Z=(z)_{n\times m}$,计算得到非零特征值 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$ 及其对应的特征向量,进而得到学术研究主体变量和研究内容变量在特征向量上的载荷矩阵[18]如下:

$$\begin{bmatrix} u_{11}\sqrt{\lambda_1} & u_{21}\sqrt{\lambda_1} & \cdots & u_{n1}\sqrt{\lambda_1} \\ u_{12}\sqrt{\lambda_2} & u_{22}\sqrt{\lambda_2} & \cdots & u_{n2}\sqrt{\lambda_2} \\ \cdots & \cdots & \cdots & \cdots \\ u_{1p}\sqrt{\lambda_p} & u_{2p}\sqrt{\lambda_p} & \cdots & u_{np}\sqrt{\lambda_p} \end{bmatrix}$$

$$\begin{bmatrix} v_{11}\sqrt{\lambda_1} & v_{21}\sqrt{\lambda_1} & \cdots & v_{n1}\sqrt{\lambda_1} \\ v_{12}\sqrt{\lambda_2} & v_{22}\sqrt{\lambda_2} & \cdots & v_{n2}\sqrt{\lambda_2} \\ \cdots & \cdots & \cdots & \cdots \\ v_{1p}\sqrt{\lambda_p} & v_{2p}\sqrt{\lambda_p} & \cdots & v_{np}\sqrt{\lambda_p} \end{bmatrix}$$

以载荷矩阵作为学术研究主体变量和研究内容变量在多维空间中的坐标,最后利用坐标计算各个变量之间的距离,得到距离矩阵 $D=(d)_{(n+m)\times(n+m)}$,用 $1-d_{ij}$ 表示主体变量或内容变量 i、j 之间的关联数值,

完成关联度的计算。

以上计算过程与经典对应分析不同之处在于:对应分析选择载荷矩阵的前二列或三列作为二维或三维空间中点的坐标,并以这些列对应的特征值和与所有特征值和的比值即累计贡献率,作为结果可信的评价指标。一般情况下,累计贡献率小于1,随着累计贡献率的降低,分析结果可信度降低。本文选择载荷矩阵的所有列作为P维空间点的坐标,累计贡献率为1。

3.2 可视化空间含义映射

应用二维空间中的网络图进行学术研究主体与研究内容间关联关系的可视化含义映射:以网络图中的节点表示学术研究主体和学术研究内容,并用节点形状加以区分,如圆形节点表示学术研究主体、矩形节点表示学术研究内容;用节点间的连线表示关联,线的粗细表示关联的强弱,关联性越大、线越粗,反之越细;节点旁的文字和数字用以标注节点所代表的学术研究主体或学术研究内容及其频率数值,连接线旁的数字用以标注连接线所表示的关联强度数值。其中频率数值和关联强度数值按照用户的需求进行显示或隐藏。

3.3 可视化算法及改进

应用复杂网络算法 Fruchterman - Reingold 进行可视化空间中网络节点的布局。Fruchterman - Reingold 算法由 Fruchterman 和 Reingold[19] 提出,简称 FR 算法。该算法建立在粒子物理理论的基础上,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。算法通过考虑原子间引力和斥力的互相作用,计算节点的速度和加速度,节点的运动规律类似原子或者行星间的运动,系统最终进入一种动态平衡状态。

在FR算法的实现上,本文做了以下改进:在计算引力时,预先设定一个关联度阈值,只有那些连接线所代表的关联度超过该阈值时,才计算连接线两端节点的引力,低于这个阈值的连接线不显示,也不计算其两端节点的引力。这样改进的益处在于:用户随时调节阈值,把那些明显的关联关系突显出来,同时加速算法本身的计算速度。

3.4 可视化图形压缩

为了突出那些关联强度显著的关系,需要对可视化网络图进行压缩,去掉那些关系不显著的连接线,识别关键信息。本文采用以下方案进行可视化网络图的压缩:

(1)应用 Pathfinder 算法,建立网络图中所有节点间最有效的连接路径。

Pathfinder 算法是美国心理学家 Schvaneveldt 等[20] 在 1989 年提出的用来分析数据相似性的一个模型,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,检查所有数据之间的关系,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信息科学技术学院 Chen[21,22] 首先使用 Pathfinder 算法实现了对超文本链接网络聚类,并在其设计开发的可视化工具 CiteSpace 中进行固化。目前,该算法在情报分析和知识可视化中已经得到广泛应用。

(2)设定关联强度阈值使可视化图形中仅显示超过设定阈值的连接线。

采用该方法进行图形压缩,方法简单,易于理解,并即时调整阈值,满足不同标准的分析需求。但每次调整阈值后,需要对网络图重新布局,增加计算时间成本和用户交互的技术复杂度。

4 实证应用

选择图书情报与数字图书馆学科 2010 年 - 2011 年间信息可视化、知识图谱相关的核心期刊论文共 342 篇,对这些论文中数量排名前 30 的机构和关键词应用本文构建的关联可视化分析方法,揭示这些研究机构之间、关键词之间、以及研究机构和关键词之间的关联关系。

图 1 为应用 Pathfinder 进行网络压缩后的关联可视化结果,图 2 - 图 4 是在图 1 的基础上分别设置关联度阈值为 0.60、0.70、0.80 的关联可视化结果。其中连接线旁的数值为关联度的小数部分。

综合图 1 - 图 4 的可视化结果可知:

(1)从机构之间的关联角度来看,中国科学院武汉文献情报中心与上海图书馆上海科学技术情报研究所之间具有较强的关联性,关联度为 0.90;南京大学历史学系与浙江树人大学图书馆之间具有较强的关联性,关联度为 0.88,南京大学历史学系与南京大学信息管理系的关联强度也达到 0.74。

(2)从关键词之间的关联角度来看,文献计量、研究前沿、CiteSpace、知识图谱、可视化图谱以文献计量学为中心形成一组关键词集合;文本挖掘、聚类分析、图书情报学、社会网络分析以多维尺度分析为中心形

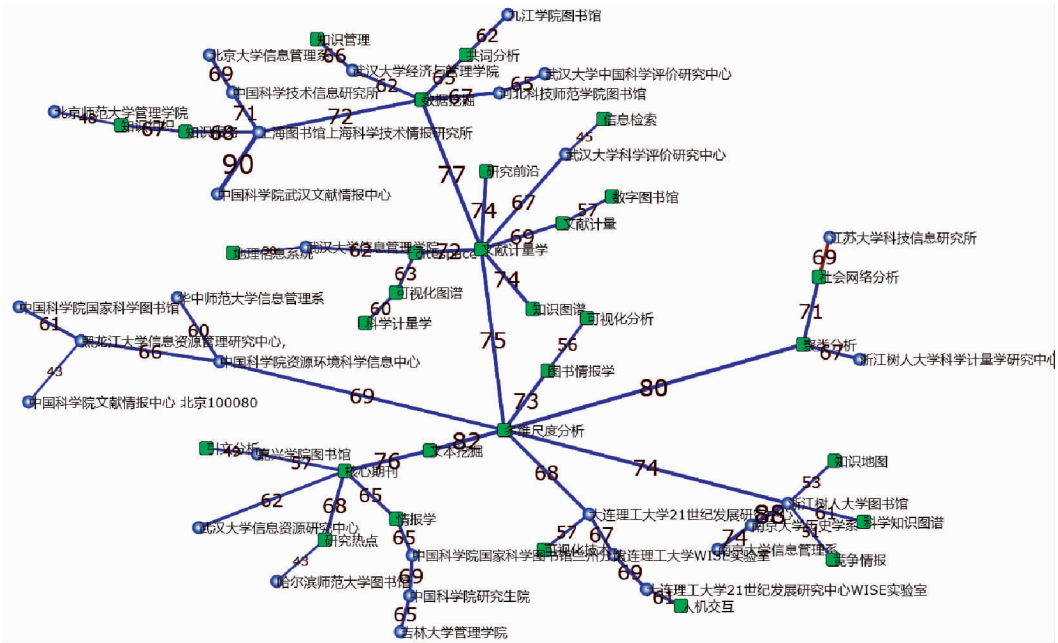

图 1　研究机构与论文关键词关联可视化(未设定关联阈值)

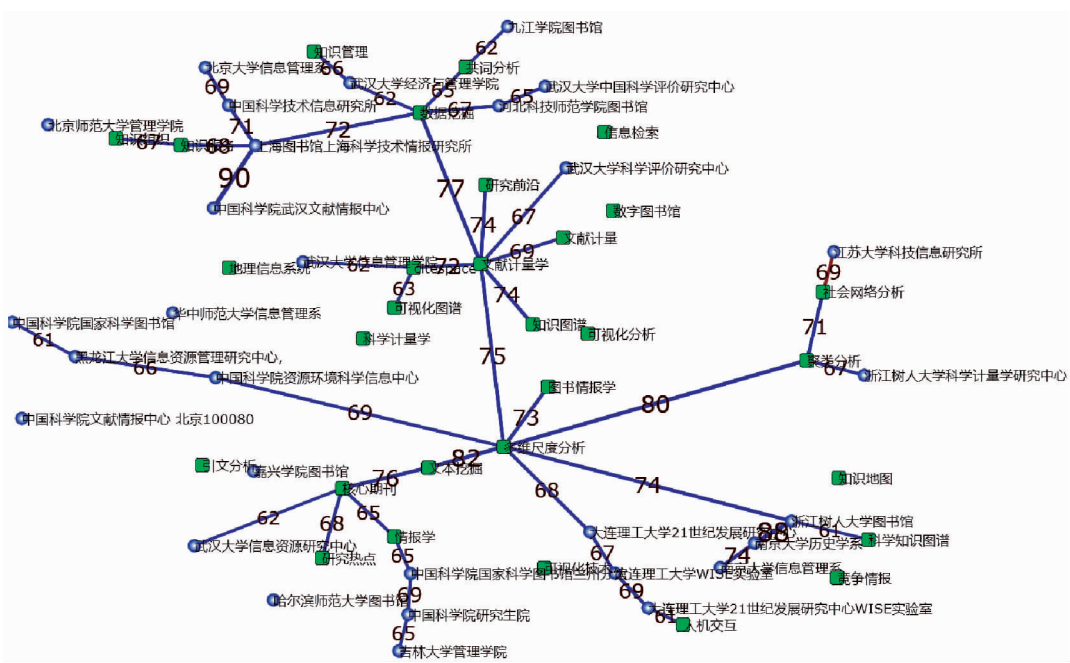

图 2　研究机构与论文关键词关联可视化(设定关联阈值为 0.60)

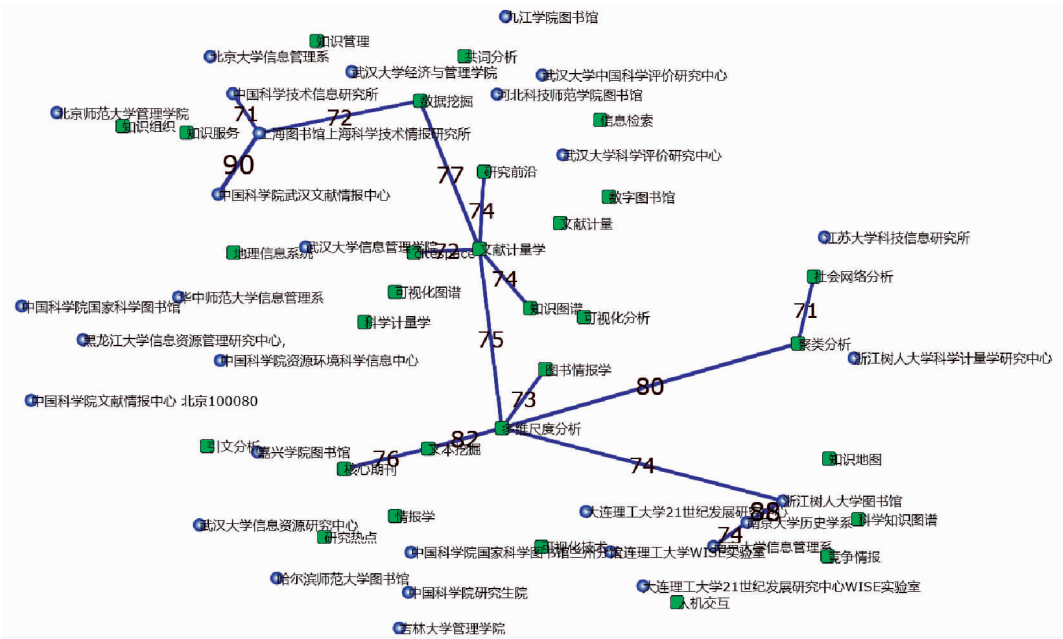

图 3　研究机构与论文关键词关联可视化(设定关联阈值为 0.70)

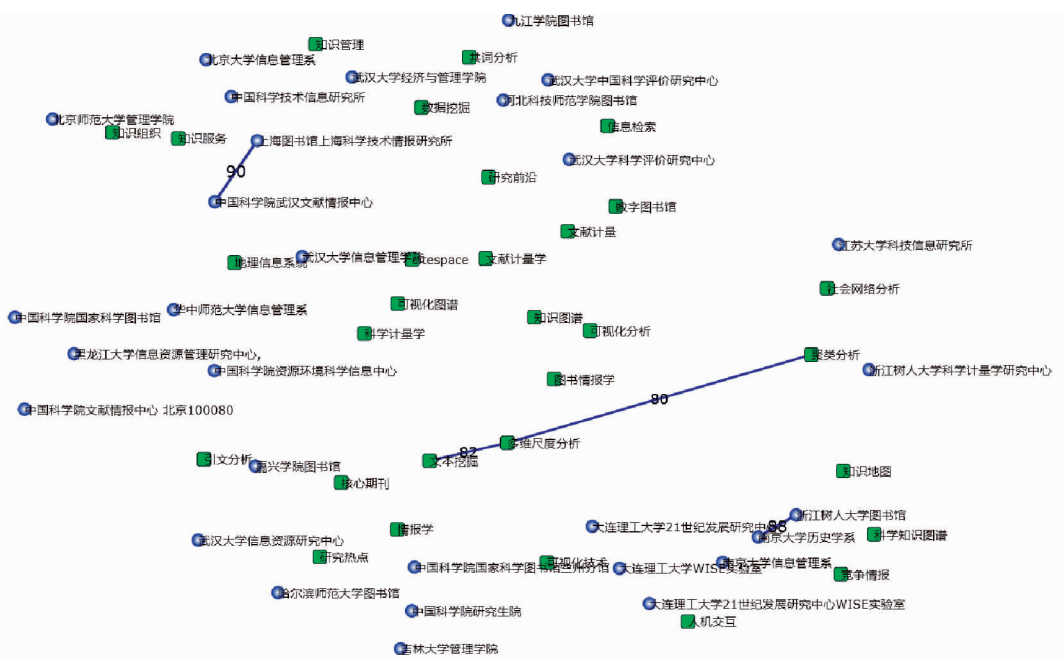

图 4　研究机构与论文关键词关联可视化(设定关联阈值为 0.80)

成一组关键词集合,关键词之间的关联强度多在 0.70 -0.80 之间。

(3)从机构和关键词之间的关联角度来看,机构与关键词之间的关联度基本在 0.70 以下,相对于机构之间的关联和关键词之间的关联要弱。分析原因主要是研究机构所使用的关键词分散,不够集中,而且每个关键词被同一机构使用的频率较小,最多的频率仅为 5,大部分为 1 或 2。

5　结　语

本文设计实现了一种学术研究主体与研究内容之间的关联关系可视化方法,将对应分析的多元统计分析方法应用于关联度的计算过程,采用网络图进行可视化呈现,以此来揭示学术研究主体之间、研究内容之间以及学术研究主体和学术研究内容之间的关联关系。如何改进距离和关联度的转化方式以更加清晰地表现关联关系,以及基于本文建立的可视化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] Zhu D H, Porter A L. Automated Extraction and Visualization of Information for Technological Intelligence and Forecasting[J]. *Technological Forecasting and Social Change*, 2002, 69(5):495 – 506.

[2] 王立学,孙杨,杨代庆. 基于引文的情报学领域主题关联特征分析[J]. *情报杂志*,2012,31(10):27 – 31. (Wang Lixue, Sun Yang, Yang Daiqing. Capturing Topics Linkage of Information Science Based on Citations[J]. *Journal of Intelligence*,2012,31(10):27 – 31.)

[3] Zha X J, Chen M H. Study on Early Warning of Competitive Technical Intelligence Based on the Patent Map[J]. *Journal of Computers*,2010,5(2):274 – 281.

[4] 王立学,冷伏海. 基于文本结构解析的动态共词方法研究[J]. *图书情报工作*,2010,54(24):37 – 40. (Wang Lixue, Leng Fuhai. Research on a Text – structure – based Dynamic Co – Word Method [J]. *Library and Information Service*, 2010, 54(24):37 – 40.)

[5] 王昊,苏新宁. 基于 CSSCI 本体的学科关联分析[J]. *现代图书情报技术*,2010(10):11 – 16. (Wang Hao, Su Xinning. Subject Association Analysis Based on CSSCI_Onto[J]. *New Technology of Library and Information Service*,2010(10):11 – 16.)

[6] 暴海龙,李金林. 专利技术关联性分析方法研究[J]. *科研管理*,2004,25(S):3 – 8. (Bao Hailong, Li Jinlin. Study of Patent Technology Association[J]. *Science Research Management*, 2004, 25(S):3 – 8.)

[7] 姚红. 基于灰色关联分析法的期刊综合评价[J]. *情报科学*,2003,21(7):730 – 734. (Yao Hong. Comprehensive Evaluation of Science and Technology Journals Based on Grey Correlation Analysis Method [J]. *Information Science*,2003,21(7):730 – 734.)

[8] 熊拥军,陈春颖. 基于关联挖掘技术的数字图书馆个性化推送服务[J]. *图书情报工作*,2010,54(1):125 – 129. (Xiong Yongjun, Chen Chunying. The Studies of Digital Library Personal Information Push Service Based on Association Rule[J]. *Library and Information Service*, 2010,54(1):125 – 129.)

[9] Doré J C, Ojasoo T, Okubo Y, et al. Correspondence Factor Analysis of the Publication Patterns of 48 Nations over the Period 1981 – 1992[J]. *Journal of the American Society for the Information of Science*, 1996,47(8):588 – 602.

[10] Doré J C, Dutheuil C, Miquel J F. Multidimensional Analysis of Trends in Patent Activity[J]. *Scientometrics*,2000,47(3):475 – 492.

[11] Bhattacharya S, Pal C, Arora J. Inside the Frontier Areas of Research in Physics: A Micro Level Analysis[J]. *Scientometrics*, 2000,47(1):131 – 142.

[12] Anuradha K T, Urs S R. Bibliometric Indicators of Indian Research Collaboration Patterns: A Correspondence Analysis[J]. *Scientometrics*,2007,71(2):179 – 189.

[13] Anuradha K T, Gopalan T K. Trend and Patterns in Explicit Organizational Knowledge: A Correspondence Analysis and Cluster Analysis[J]. *The International Information & Library Review*, 2007,39(3 – 4):247 – 259.

[14] Iribarren – Maestro I, Lascurain – Sánchez M L, Sanz – Casado E. Are Multi – authorship and Visibility Related? Study of Ten Research Areas at Carlos III University of Madrid[J]. *Scientometrics*, 2009,79(1):191 – 200.

[15] 刘玉琴. 基于专利检索与专利分析的技术创新管理方法研究[D]. 北京:北京理工大学,2008. (Liu Yuqin. Study on Methods of Technological Innovation Management Based on Patent Retrieval and Patent Analysis[D]. Beijing: Beijing Institute of Technology, 2008.)

[16] VantagePoint [EB/OL]. [2012 – 11 – 12]. http://thevantagepoint.com/.

[17] Thomson Data Analyzer[EB/OL]. [2012 – 11 – 12]. http://www.thomsonscientific.com.cn/media/tda.Pdf.

[18] 于秀林,任雪松. 多元统计分析[M]. 北京:中国统计出版社,1999:199 – 201. (Yu Xiulin, Ren Xuesong. Multivariate Statistical Analysis[M]. Beijing: China Statistics Press,1999:199 – 201.)

[19] Fruchterman T M J, Reingold E M. Graph Drawing by Force Directed Placement[J]. *Software Practice and Experience*, 1991, 21(11):1129 – 1164.

[20] 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.

[21] Chen C M. Generalised Similarity Analysis and Pathfinder Network Scaling [J]. *Interacting with Computers*,1998,10(2):107 – 128.

[22] Chen C M. CiteSpace Ⅱ: 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.

(作者 E – mail:liuyuqin2004@126.com)