信息可视化在医学文献分析中的初步应用理论研究*

王敏 张燕舞 张玢 李海存 刘晓婷 许培扬 肖永红

（中国医学科学院医学信息研究所 北京 100020）

（中国科学院国家科学图书馆 北京 100190）

〔摘要〕 信息可视化是情报学重要研究领域之一，借助文献分析、可视化工具对大量文献数据信息绘制科学知识图谱，为清晰、准确地揭示知识领域结构，发现科技研究热点、研究前沿提供了新的手段。在调研国内外信息分析可视化研究的基础上，总结信息分析可视化的常用方法、工具及其主要应用领域，为在医学领域开展信息分析可视化应用研究奠定基础。

〔关键词〕 信息可视化；知识图谱；内容分析；引文分析；社会网络分析

Theoretical Research on Primary Application of Information Visualization in Medical Literature Analysis *WANG Min, ZHANG Yan-wu, ZHANG Bin, LI Hai-cun, LIU Xiao-ting, XU Pei-yang, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020*, China; *XIAO Yong-hong, National Science Library, Chinese Academy of Sciences, Beijing 100190*, China

〔**Abstract**〕 Information visualization is one of the important subfields of information science. Mapping scientific knowledge domain by applying literature analysis and visualization tools on a lot of literature data information provides a new method that attempts to reveal the intellectual structure, track the dynamics, and discovers the potential new areas of research. Based on the investigation of information visualization research both at home and abroad, the paper presents an overview of the commonly used methods, tools and main application fields, which lay a foundation for carrying out application research of information analysis visualization in medical area.

〔**Keywords**〕 Information visualization; Knowledge domain; Content analysis; Citation analysis; Social network analysis

生物医学作为当今科技发展最为迅速的领域之一，生物医学信息资源迅猛增加。如何利用科学方法、先进技术对其进行快速、有效挖掘，深度揭示医学科技研究热点、研究前沿，辅助科研人员确定科研选题方向、把握科技研究现状，为科研管理人员、科技政策人员制定科技决策提供支持服务，成为医学信息分析人员共同关注的问题，而分析结果的有效揭示、呈现是其中一个重要研究内容。

信息可视化是将抽象数据用可视的形式表示出来，可用于知识发现、决策制定、信息检索、信息系统界面设计、数字图书馆、文献信息分析等领域。目前信息可视化已成为近几年情报学领域研究的新热点，在国外研究相当活跃，但在我国研究尚处于起步阶段。本文拟通过总结国内外信息分析可视化的方法、工具及其主要应用领域，为我国在医

〔收稿日期〕 2009-01-19

〔作者简介〕 王敏，馆员，发表论文 8 篇。

〔基金项目〕 中国医学科学院医学信息研究所中央级公益性基本科研业务费专项资助课题“信息可视化在医学信息分析中的应用研究”（项目编号：08R0129）。

学领域开展信息分析可视化应用研究奠定基础。

1 信息分析可视化研究概述

1.1 国内外信息分析可视化研究现状

（1）国外 文献信息可视化是多学科交叉、前沿性的研究工作。美国科学情报研究所和美国德雷克塞尔大学一直是信息可视化研究最活跃的阵地，从 20 世纪 60 年代至今，已涌现了一批信息分析领域专家，如加菲尔德（Garfield）、格里菲思（Griffith）、怀特（White）、麦肯（Macain）和陈超美（Chen CM）等，从信息分析可视化相关方法的研究到相关工具的开发、应用都取得了突出的成就。

（2）国内 国内近年来也陆续开展了信息可视化、知识图谱研究，大连理工大学刘则渊教授带领他的团队在国内率先发表了一批科学知识图谱方法及其应用的论文和专著，北京理工大学朱东华教授带领他的团队开展了多项科技文献信息可视化数据挖掘方面的研究，这些研究为国内信息可视化在文献信息分析、科学计量分析方面的研究奠定了基础。

（3）国内外应用领域 对近 5 年 Web of Science、PubMed 收录的信息分析可视化文献分析显示，国外信息分析可视化应用领域主要有图书馆学、信息科学、管理学等，在医学领域应用较少。对国内中国知网数据库收录的文献进行可视化分析显示，信息分析可视化、知识图谱应用领域主要局限于管理学、科学学等，在医学领域应用研究甚少。

1.2 信息分析可视化的特点

本课题研究的信息可视化主要着重于分析结果的可视化，而不是分析过程的可视化。从分析结果角度，可视化技术主要分为两大部分：传统的统计数据可视化和新的信息可视化技术[1]。

（1）传统的统计数据可视化 主要采用各种图表，如柱状图、折线图、XY 散点图、直方图、雷达图、气泡图等。主要存在两个问题：一是统计数据图形分类基于图形外观形态，而不是基于内在统计数据表现机理，以致可视化图形选择带有一定的盲目性；二是统计数据可视化图形种类较少，较难满足多维统计数据表现和分析的需求。

（2）新的信息可视化技术 主要采用几何投影技术、面向像素技术和分层技术，能较好解决上述问题。显示的图形、图像主要有二维、三维、多维信息，以及时间序列信息可视化、层次信息可视化和网络信息可视化等[2]。

（3）将信息可视化技术应用于文献分析有两个比较突出的优点 一是直观可视性，提供了直观理解大量数据的途径，通过可视化辅助用户辨别重要信息。如利用 CiteSpace 绘制知识图谱，辅助用户寻找科技发展中的关键点、把握某一学科或知识域的前沿领域[3]；利用专利分析工具 Aureka 生成 Thememap 图，纵览技术领域布局、揭示技术/竞争对手间的关系等；二是多维性，信息分析可视化可将三维以及多维用直观的方式显现出来，从多方面揭示事物间的联系。如 Vxinsight 通过虚拟地形图（Landscape）来模拟聚类信息，可用于扫描数据集和揭示数据间的相似关系。

2 信息分析可视化常用方法及可视化显示

2.1 基本统计分析

基本统计分析是利用统计学方法对文献进行统计分析，以数据来描述和揭示文献的数量特征和变化规律，从而达到一定研究目的的一种研究分析方法[4]，分析结果通常以列表、直方图表等传统的统计数据可视化形式展现。

2.2 内容分析[5]

内容分析是通过对研究对象的题名、关键词、摘要等进行分析，来挖掘科技领域研究热点、跟踪科技领域前沿的方法。主要包括共词分析、共引分析、聚类分析、多维尺度分析等。

（1）共词分析 是对一组词（从文献题名、关键词、摘要或正文中抽取）两两统计它们在同一篇文献中共同出现的次数，以此为基础对其进行聚类分析来反映词间亲疏关系，以及这些词代表的学科或主题的结构与变化[6]。共词分析主要用于揭示知识领域结构、映射知识领域发展趋势，可视化图谱主要包括战略坐标图、节点——链接图和 MDS 图 3 种[7]。

（2）共引分析 是指两篇文献同时被后来的文献所引用，其实质是将一组具有共引关系的文献作为分析对象，综合利用数学、统计学和逻辑分析方法，把对象间错综复杂的共引关系量化、抽象并简单表达的过程，用于揭示文献间的关联度和内容的相似性。共引分析主要用于探测和分析学科研究热点、研究前沿，目前应用较多的共引分析可视化图谱主要基于 CiteSpace 生成共引网络图，辅助研究者较直观地辨识学科前沿演化路径及其经典基础文献[8]。

（3）聚类分析 是最常用的多元统计分析方法之

一，其目的是把分类对象按一定规则分成组或类，这些组或类不是事先给定的，而是根据数据特征而定的。聚类分析可视化是将聚类分析方法与共词分析、共引分析相结合，对高频词、高被引文献进行聚类分析，用于反映学科主题领域结构、研究热点及研究前沿。可视化图谱主要采用 SPSS 软件生成树状图。

（4）多维尺度分析（Multidimensional Scaling，简称 MDS） 是指通过某种非线性变换，把高维空间的数据转换成低维空间中的数据，变换后的数据仍能近似保持原数据几何关系的一种技术，即通过低维空间（通常是二维空间）展示分析对象间的联系，并利用平面距离来反映对象间的相似程度。可视化图谱主要采用 SPSS 软件生成 MDS 图。由于 MDS 图较难确定点群的边界和数目，因此常将 MDS 与聚类分析或主成分分析（因子分析）配合起来展示多维共现矩阵在二维空间的关系[9]，以便更清晰准确的揭示学科研究主题领域。

2.3 引文分析

引文分析一般采用文献计量和统计的方法，以期刊、论文、专著等为研究对象，对其引用与被引用规律进行分析，揭示它们所蕴含的内在联系。引文分析，可用于揭示科学发展规律，评价科学现象、监测研究热点[10]。引文分析应用研究主要有引文数量分析、引文网络时序分析、共引分析和耦合分析，引文分析图谱主要包括引文网络时序图、共引网络图谱和时间线视图。

（1）引文网络时序图 在时间纵轴上对文献先后引用关系及其重要性进行分析，揭示某个研究主题论文源流、最初著者及其发展脉络，并从中探讨科技发展历程和研究规律。2001 年加菲尔德等研制开发了 Histcite 软件，该软件基于 SCI 数据，根据文献间的互相引用关系，按照年份顺序生成针对某一研究主题的引文编年图（Histography）[11]。

（2）共引网络图谱 依据文献分析单元可分为文献共引网络图谱、期刊共引网络图谱、作者共引网络图谱 3 大类，其原理类似。文献共引网络图谱主要基于 CiteSpace 软件，探测和分析学科研究前沿随时间变化趋势以及研究前沿与其知识基础之间的关系[12]；期刊共引、作者共引可采用 Ucinet、SPSS 的聚类分析功能，绘制学科领域主流期刊、学术群体，揭示期刊、作者间的相互依赖和交叉关系。

（3）时间线视图 在时间横轴上对文献间先后引用关系及其重要性进行分析，利用文献耦合方法对其进行聚类分析，根据生成的文献簇在时间线视图上的分布和文献簇间引用关系揭示研究前沿。时间线视图可采用 Morris 等开发的 DIVA 软件实现[13]。

2.4 社会网络分析

社会网络分析（Social Network Analysis，简称 SNA）是社会科学和行为科学中一种独特的研究视角，注重单元间相互关系，其实质是用点和线表达网络。SNA 通过计算关联主题数量方法识别主题网络中的核心主题（核心点）和次要主题（非核心点），关联主题数量最多的为核心主题，其他为次要主题，其顺序由测地线距离决定[14]。目前 SNA 主要用于学科主题、作者、机构、国家间关系的聚类和可视化展示，SNA 图谱可采用 Pajek 或 Ucinet 实现。

3 信息分析可视化常用工具

3.1 种类

目前国际上信息可视化工具主要分为可视化开发工具和可视化分析工具两大类，开发工具比较典型的有 Piccolo、Prefuse[15]，分析工具典型的有 CiteSpace、Vxinsight 等。由于应用信息可视化开发工具需要一定的编程基础，因此本文着重对信息可视化分析工具进行介绍，并将其分为基于文献计量的分析工具和基于社会网络的分析工具两大类。

3.2 基于文献计量的分析工具

基于文献计量的分析工具，国外开发较多，国内尚没有商业化的分析工具。通过文献调研、专家咨询，国外信息分析可视化研究应用较多的主要有 Thomson Data Analyzer（简称 TDA），CiteSpace，Histcite，Vxinsight，DIVA，由于 DIVA 需要 Metalab 编程工具支持，在此重点对前 4 种工具基本概况及其在学科情报研究服务中可实现的功能进行总结，见表 1，表 2。

表 1 **TDA，CiteSpace，Histcite，Vxinsight 的基本概况比较**

项目	TDA	CiteSpace	Histcite	Vxinsight
最新版本*	2.1	2.2 R3	8.12.16	2.145
付费类型	商业	免费	商业	免费
面向对象	科技文献	科技文献	科技文献	科技文献、基因组、蛋白组研究
主要数据源	Derwent Innovations Index, Web of Science, Medline, Biosis Previews, Inspec 和 Current Contents Connect 等，兼容 Excel, Delphion, Aureka 等	Web of Science, Pubmed, Citeseer	Web of Science	MS Access Database
数据清理	有	无	无	无
分析方法	基本数量统计、共现分析	共现分析	基本数量统计、引文分析	聚类分析
结果生成	统计图表、共现矩阵、节点链接图、技术报告	节点链接图、引文网络图谱	统计表、引文时序网络分析	虚拟地形图
图谱维度	二维	二维	二维	二维，三维

注：最新版本截至时间为 2009 年 10 月 12 日。

表 2 **TDA，CiteSpace，Histcite，Vxinsight 学科情报服务功能比较**

项目	学科情报分析任务[16]	TDA	CiteSpace	Histcite	Vxinsight
调研科学研究领域的基本情况	学科发展趋势分析	√	√	√	-
	确定国内重点学科带头人	√	√	√	-
	科学技术间的内在联系	√	√	-	-
为选择和确定科研项目提供学科战略情报服务	国际学科布局	√	√	-	√
	学科热点分析	√	√	√	√
	国外著名学术机构科学家科研动向	√	√	√	√
	国外科研项目的研究现状	√	√	-	-
	国内外科研成果差距领域	√	√	-	-
国内外学科发展的比较优势及相对影响力对比	机构或国家科技竞争力对比	√	-	-	-
	国际科研合作分析	√	√	-	-
揭示学科内部的知识结构类问题	知识结构图	√	√	√	√

3.3 基于社会网络的分析工具

SNA 是研究社会关系的一种新兴的研究方法，基于社会网络的文献信息分析与可视化工具在分析学科研究领域、科研合作网络中起着重要作用。目前国内外应用较多的社会网络分析工具主要有 Pajek 和 Ucinet，具体比较，见表 3。

表 3 **Pajek 与 Ucinet 的比较**

比较内容	项目	Pajek	Ucinet
版本		v1.26（2009-9-3）	v6.242（2009-10-8）
付费类型		免费	商业
软件性质		单机版	单机版，集成 Pajek、NetDraw 和三维展示分析软件 Mage
适用对象		大型数据网络可视化	综合
数据格式	类型	完全数据，从属数据，大型网络数据	完全数据，自我为中心的数据，大型网络数据，从属数据
	输入格式	矩阵，链接/节点，节点	矩阵，链接/节点，节点
	缺省值	有	有
功能	可视化	有	有
	分析类型	描述性的，结构和特定区域，角色和位置，二元组合三元组方法	描述性的，结构和特定区域，角色和位置，二元组合三元组方法，统计

4 信息分析可视化的主要应用

4.1 学科热点分析

学科热点分析在学术研究中起着重要作用，它可以为管理人员科学决策提供参考，为学科研究人员确定研究方向和研究内容提供情报依据[17]。国内外学科热点研究主要着重于学科热点主题演化分析[18]和学科热点跟踪研究[6]，且均采用定性分析、定量分析相结合的方式开展。定性分析法是由领域专家依靠专业知识、研究经验和综合分析判断的能力，对学科研究热点主题随时间的发展变化情况做出判断。定量分析是通过统计方法量化不同时期的主题网络的演化情况。目前，开展学科热点分析常用定量方法主要基于内容分析，如共词分析、共引分析、聚类分析等，可视化图谱展示主要有战略坐标图、节点——链接图、MDS 图、共引网络图谱等。

4.2 研究前沿分析

在全球科技竞争日益加剧的今天，如何能够科学、准确地把握科学研究前沿已成为各个科学技术领域专家开展技术预测关注的焦点，更成为各国政府制定科技发展战略时面临的一大问题。目前，研究前沿还没有统一的认识和定义，一般认为研究前沿是科学研究中最先进、最新、最有发展潜力的研究主题或研究领域。国内外科技前沿分析主要采用共词分析[19]、共引分析[20]和耦合分析[21]等，可视化图谱主要有共引网络图谱、时间线视图等。

5 结语

将信息可视化技术应用于医学领域文献分析，能够深层次挖掘医学领域研究主题内在关联，以可视化形式传递和呈现隐性信息，更直观有效揭示医学文献分析结果，为把握医学领域发展前沿和研究热点提供了新的手段。因此，信息分析可视化在医学文献分析中具有广阔的应用前景，信息分析可视化将成为深入开展医学文献分析的重要技术支撑。

在课题组开展的肝移植、肝干细胞信息分析可视化应用中发现，信息分析可视化的结果尚存在一些不足，如易读性差，图谱中节点——链接较多，存在节点重叠现象，或是图谱中缺少标识，用户较难理解各符号含义。正如国际情报学专家 Havard 教授指出的，可视化图谱目的在于提高分析结果可读性，信息可视化研究人员应注重提高可视化工具的用户友好性、图谱可读性、合理性，能让阅读者较清晰地读懂图和其中的关系，在信息分析可视化图谱展示方面应简单化，不能为了作图而作图[22]。

此外，信息分析可视化的应用尚存在较多问题有待进一步研究，如不同算法对可视化结果的影响、可视化分析过程中的人机交互问题、可视化技术结合文献计量方法的深层次应用等，都需要继续实践和探讨。

参考文献

1 康宇航. 一种基于共现分析的科技跟踪方法研究〔D〕. 大连：大连理工大学，2008.

2 李纲，郑重. 信息可视化应用研究进展〔J〕. 图书情报知识，2008，(4)：36－40.

3 陈超美，陈悦，侯剑华. CiteSpaceII：科学文献中新趋势与新动态的识别与可视化〔J〕. 情报学报，2009，28(3)：401－421.

4 宋巧枝，方曙. 基于文献统计分析法的专利计量分析研究〔J〕. 现代情报，2008，(2)：125－129.

5 Chen，H，Fuller，SS，Friedman，C，et al. Medical Informatics：knowledge management and data mining in biomedicine〔M〕. Springer，2005：41－42.

6 He，Q. Knowledge Discovery Through Co－word Analysis〔J〕. Library Trends，1999，48 (1)：131－159.

7 赵凡. 基于共词分析的学科主题动态跟踪相似方法研究〔D〕. 北京：中国科学院文献情报中心，2008.

8 侯剑华，陈悦. 战略管理学前沿演进可视化研究〔J〕. 科学学研究，2007，25 (增刊)：15－21.

9 Yulan He，Siu Cheung Hui. Mining a Web Citation Database for Author Co－citation Analysis〔J〕. Information Processing and Management，2002，(38)：491－508.

10 庞龙. 科学引文分析的科学评价功能和意义〔D〕. 太原：山西大学，2006.

(下转第 49 页)