

基于“全景 + 细节”策略的学术引文信息可视化研究*

靖培栋 邱 威 牟晟斓

摘 要 在对国内外引文信息可视化系统进行分析的基础上,基于“全景 + 细节”策略,运用可视化方法对从 CNKI 获取的原始引文信息进行研究。通过可视化的引文网络、分支信息、分支网络,可全方位、多角度查看引文情况;通过基于引文耦合的聚类信息可视化,可以发现研究热点;通过文献及作者耦合可视化,可以了解作者的研究兴趣的关联;最后对洛特卡定律进行了可视表示。未来的研究应注重扩展信息可视化系统的功能,改进信息可视化。图 7。参考文献 13。

关键词 信息可视化 引文分析 “全景 + 细节”

分类号 G350.7

ABSTRACT This paper conducts the visualization research on academic citation information with *overview + detail* strategy. Through visualized citation networks, components information and components network, we can clearly examine citation status from multiple views. Through the visualization of information clusters based on citation coupling, we can discover academic hotspots. Through visualization of documents coupling and authors coupling, we can discover the relationship of authors' research interests. At last, it visually gives the representation of Lotka's law. 7 figs. 13 refs.

KEY WORDS Information visualization. Citation analysis. *Overview + detail.*

CLASS NUMBER G350.7

1 前言

引文分析是一种重要的文献计量分析方法,引文中往往包含许多规律与有价值的信息,这也是人们做引文分析研究的原因。近几年在国内外出现了一些引文信息可视化方面的系统。美国 Drexel 大学 Chaomei Chen (陈超美)开发了用于可视化文献资源关系的系统 CiteSpace^[1],这一系统在国内应用较广泛,许多引文分析学者使用这一软件,分析某些领域的研究状况。引文可视化分析工具 HistCite^[2],是通过生成文献编年图表来实现领域分析的功能,国内学者对这一系统进行了介绍与评价^[3]。除此以外,美国 Drexel 大学的学者还开发了基

于作者共引分析(ACA)的 AuthorLink、基于词共现的 ConceptLink、集成多种类型的索引和多种视图的 PNASLink 三种基于引文的信息可视化系统^[4]。国内学者孙巍和张学福对这些系统进行了对比分析^[5]。瑞典学者也开发了科学引文网络可视化系统 CiteWiz^[6]。张秀梅和吴巍基于万方数据生物医学期刊,对其作者在合作研究中形成的复杂网络进行了可视化研究^[7]。李运景和侯汉清则对引文分析可视化技术进行了分析,指出了其中存在的问题^[8]。

以上诸多系统中,HistCite 系统的可视化模式较符合文献计量学学者的习惯,但其中的可视化功能较弱。CiteSpace 及 CiteViz 系统中的可视化模式,与文献计量学者特别是中国的文献计量

* 本文系国家社科基金资助项目“信息可视化研究”(项目编号:02BTQ019)研究成果之一。

学者的研究习惯有一定差距。而 AuthorLink、ConceptLink 及 PNASLink 系统的功能较单一。而且,以上系统中各种关系是基于共引 (Co-Citation) 建立的,而笔者认为基于耦合 (Coupling) 的聚类,更能够及时揭示学术前沿与热点。

基于以上考虑,本研究运用全景+细节 (overview + detail) 策略^[9],使用可视化方法对学术引文信息进行研究。通过可视化给出引文网络、分支信息、分支网络等,从而全方位、多角度察看引文情况;通过基于引文耦合的聚类信息可视化,可以发现研究热点;通过作者耦合可视化,可以了解作者的研究兴趣的关联;通过可视化,还可以看到洛特卡定律的成功表示。这种策略的核心思想主要是通过呈现多重视图使得用户能够同时从多个视角来查看内容,既能看到全景又能看到指定的细节。比如,在一个窗口中显示全景视图,在另一个窗口中显示指定细节视图,这样用户便可以在对全部内容有一个整体认识的同时,还能关注自己感兴趣部分的细节。这样的策略通常见于一些地图或图像浏览软件中。

2 数据及其存储

本研究通过中国学术期刊网 (CNKI) 选择获取发表在《中国图书馆学报》、《情报学报》及

《图书情报工作》三种核心期刊上图书情报领域的论文作为原始数据,年代为 1999—2008 年,论文题目中包含“检索”关键词,称其为原始文献,共 366 篇,记录下它们之间的引用关系。再加上这些论文中部分论文的引文,共计 929 篇。再利用 SQLServer 2000 对这些论文及其引用关系进行存储,形成引文数据库。引文数据库主要分为两部分:第一部分为论文自身的基本信息,包括论文的题名、作者、文献出处、文献的发表日期等,并且为每篇论文赋予一个唯一的 ID 号。其中,在录入“作者”这一信息时,如果文献的作者为多人,则依次将作者姓名输入数据库,作者姓名之间用逗号间隔;第二部分为论文之间的引用关系,使用引证文献的 ID 和其对应参考文献的 ID 表示。

3 引文网络的可视化

3.1 整体引文网络

对引文网络^{[10]316}进行可视化,本研究主要采用引文编年图,基于引文的时序性,将整个引文网络的引用关系分为自上而下、自左向右的引用表现,清楚地表现出科学文献之间的交流情况。图 1 呈现的是原始文献间的引文网络。

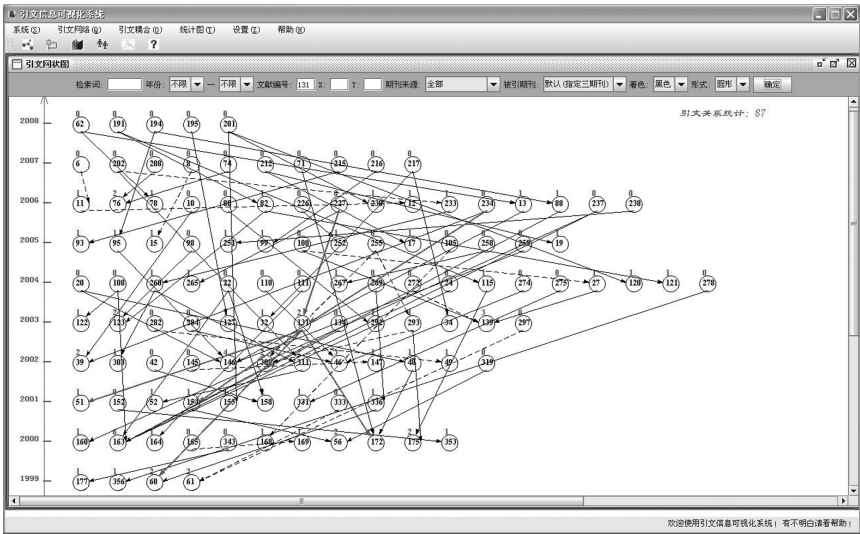


图 1 原始文献间的引文网络

图中的小圆代表论文,其中的数字是论文编号,小圆左上方的数字是该论文的被引次数,黑色实线表示两论文间的引用关系,黑色虚线表示的是自引,变粗的线表示的是特定文献的引用和被引用情况(通过输入文献编号产生),图1中显示了编号为131的文献的引用和被引用情况。在本系统中可以对关键词、发表时间、期刊来源、文献编号进行限制,以考察不同主题、不同时期、不同期刊来源的引文网络,从而可以观察学科的发展特点。通过对引文网络的整体观察,我们可以发现被引次数高的论文。如图1中编号为163的论文的被引次数为6次,说明这篇论文有较高的关注度。

3.2 分支信息呈现

为了清楚地观察与某篇论文相关的引文网络,呈现出该论文在整个研究过程中的作用,所有与某篇文献有直接或间接引用关系的文献构

成一个文献集(我们称为分支)。其实整个引文网络就是由若干个这种分支组成的,这些分支之间没有引文关系,可以说每个分支代表了不同的研究方向。图2是分支信息汇总,在图2的生成过程中,本研究使用了迭代算法。从引文数据库中的一篇文献A开始,搜索出与该文献有引用关系的文献,A与这些文献形成集合S;随后再搜出与文献集S中文献有引用关系的文献,将其合并到文献集S。依次类推,直到搜索不出新的文献,就形成这一分支中的所有文献。再从S以外的一个文献搜索,就形成新的分支。最终把整个引文网络分解成若干分支。

通过分支信息我们可以清楚地知道该引文网络的分支个数以及各分支中的文献个数。在图2中,以一条垂直的竖线为分割线,左侧的数字表示对应的右侧分支由多少篇文献构成。同时可以进一步观察分支的详细引用关系,如输入文献编号62可得如图3的分支。

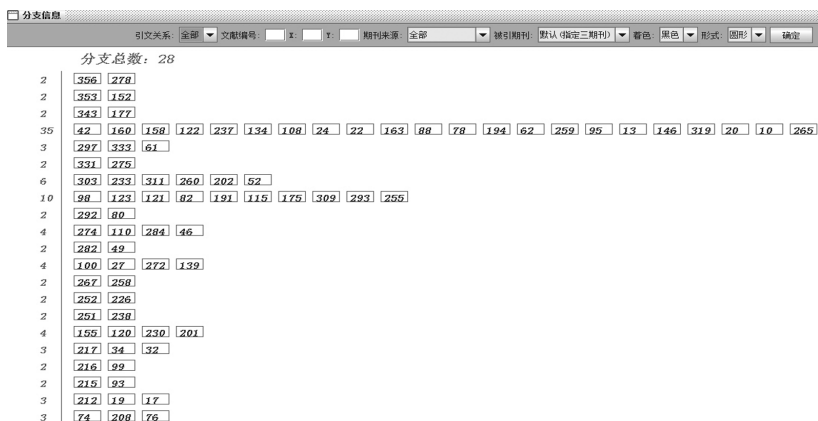


图2 分支信息

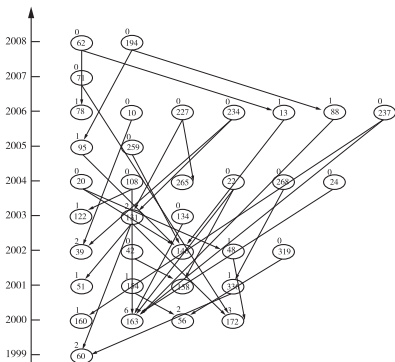


图3 分支网络

4 引文耦合可视化

4.1 引文耦合

所谓文献耦合,就是如果A和B两篇文献共同引用了一篇或多篇参考文献,那么文献A和B就具有引文上的耦合关系^{[10]398[11]}。具有耦合关系的文献必然在学科内容上存在某种联系或相关性,其相关性的强弱可以用“耦合强度”来衡量。耦合强度是根据文献共引的参考文献的篇数来计算的。耦合强度越高,文献间

的联系越紧密。引文耦合分析也称文献耦合分析,它以被引文献作为联系媒介,把文献联系起来。可以利用文献间的这种耦合关系,对文献聚类,形成文献簇。对于簇中的每一篇文献,簇中一定存在另一篇文献与其有耦合关系。基于聚类的文献簇在一定程度上能够揭示学科的发展情况和学科的前沿与热点^[12]。

实际上,“耦合”的概念不仅仅局限于文献之间,从文献间的耦合关系可以进一步引申到作者之间的耦合关系,即用文献的耦合定义作者的耦合,作者耦合反映了作者之间研究兴趣

的关联。这种分析方法在图书情报学、科学学和人才学领域有广泛的应用。

4.2 引文耦合可视化实现

4.2.1 基于耦合的聚类可视化

首先对引文数据库中的引文关系进行分析,找出所有具有耦合关系的文献对,将这些文献对作比较,若交集不空,则将其合并为文献集,如此反复直到所有文献集之间的交集为空,则形成聚类,聚类结果见图4。

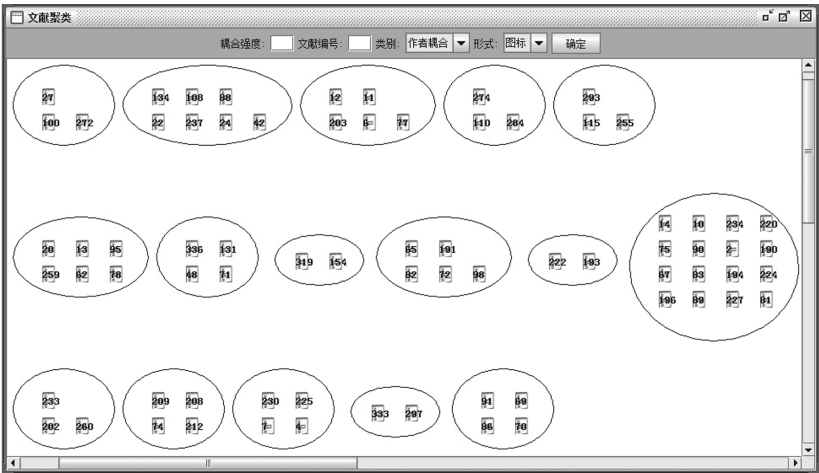


图4 文献聚类

图4中每个椭圆表示一个文献簇,椭圆内的每个图标表示一篇文献,并按一定的方式进行排列,图标上的数字是它的ID,椭圆大小可随文献簇的大小自动调整。系统提供了三种表示文献的方式:图标、圆点、编号。可以输入耦合强度来查看不同耦合强度下文献耦合的结果。由图4可知,对现有引文数据库中的文献进行耦合处理后,生成了16个文献簇。这些文献簇在一定程度上揭示了这一学科领域中不同的研究方向和研究主题,而文献簇的大小还可以帮助用户发现研究热点。如图4中最大的文献簇主题不太明确,将耦合强度调整为2后得到6个簇,其中一个为此最大簇的子集,主题是关于信息检索可视化及相关问题研究。第二大的文献簇是关于知识检索研究,第三大的文献簇是关

于语义检索研究,耦合强度为2的簇中还有一个是关于XML检索的。这说明这些研究点应该是信息检索领域中的研究热点。

4.2.2 文献耦合可视化

在系统中可以输入任一文献的ID,来查看包含此文献的簇中的文献耦合或是作者耦合情况。图5显示了当输入文献ID为44时的情形,每篇文献用一个矩形表示,矩形的左上方处显示文献的ID,矩形的中间显示文献的篇名,但如果篇名过长,则仅显示篇名的一部分,具有耦合关系的文献之间用直线或曲线连接,并用线条的粗细来表示文献间的耦合强度,线条越粗说明文献间的耦合强度越大,即文献间的内容相关度越高。

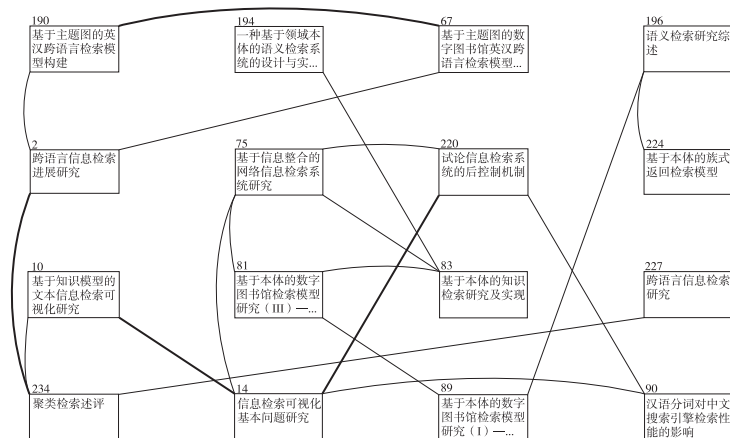


图5 文献耦合

4.2.3 作者耦合可视化

系统还可以展示作者耦合,如图6所示,用直线或曲线连接具有耦合关系的作者;同样,线条的粗细表示作者之间的耦合强度,线条越粗说明作者之间的耦合强度越大。图6只展示了文献的第一作者之间的耦合关系,通过作者之间的耦合关系,可以了解到都有哪些作者之间拥有共同的研究兴趣,或是发现与自己所关注的作者有着相关研究兴趣的其他作者,通过输入作者姓名可以显示与其有耦合关系的作者视图。而对于作者本身而言,还可以发现与自己有共同研究兴趣的作者,以便学者之间的学术交流与合作。

5 洛特卡定律的可视化表示

由于从系统中可以统计出作者发表论文数量的信息,笔者顺便对洛特卡定律^{[10]167}进行了验证。根据最小二乘法,得出了引文数据库中作者频率与论文数量之间的关系,生成了图7。算出洛特卡定律中 x 的指数 $D=2.839$, D 是洛特卡定律的核心所在,一般认为 D 在 $1.2\sim 3.5$ 之间^{[10]169[13]}。由 D 的分形意义, D 值小,说明著者分布较为集中,不平均程度高,或这一领域著者著述能力悬殊比较大,论文主要集中于高产著者。这个结果和我们的统计数据相吻合,发表一篇论文的作者为 89.32% ,这说明在我们所选择的专题“检索”中高产作者很少。

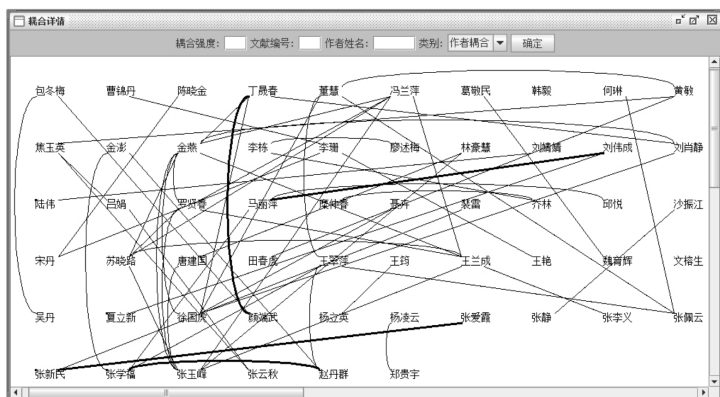


图6 作者耦合

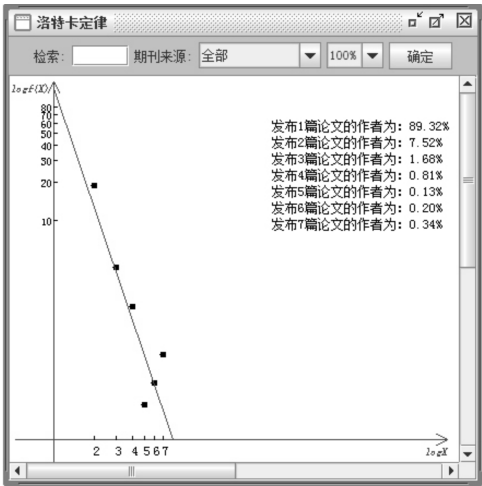


图7 洛特卡分布

6 结语

以上是本引文信息可视化系统的主要功能。通过对引文信息可视化系统的使用,我们能够深刻体会信息可视化的作用,即可以全方位、多角度查看引文情况。具体来说,通过聚类信息可视化,发现了研究热点;通过作者耦合可视化,了解到了作者的研究兴趣的关联,看到了洛特卡定律的成功验证。系统使用的信息量有限,因此得到的结果也许有些偏差,主要是为引文分析提供一种新的研究方法。系统的功能有待扩展,可视化方式需要改进,这是未来图书馆情报学信息可视化研究的方向。

参考文献:

[1] 陈超美. CiteSpace II: 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 (3): 401-421.

[2] Garfield E, Paris S, Stock W. HistCiteTM: A software tool for informetric analysis of citation linkage[J]. Information, 2006 (8): 391-400.

[3] 李运景,侯汉清,裴新涌. 引文编年可视化软件 HistCite 介绍与评价[J]. 图书情报工作, 2006 (12): 135-138.

[4] 林夏. 信息可视化与内容描述[J]. 现代图书情报技术, 2004 (10): 3-13.

[5] 孙巍,张学福. 基于引文的信息检索可视化相关系统比较分析[J]. 情报理论与实践, 2008 (4): 598-601.

[6] Elmqvist N, Tsigas P. CiteWiz: A tool for the visualization of scientific citation networks[J]. Information Visualization 2007 (6): 215-232.

[7] 张秀梅,吴巍. 科研合作网络的可视化及其在文献检索服务中的应用[J]. 情报学报, 2006 (1): 9-15.

[8] 李运景,侯汉清. 引文分析可视化研究[J]. 情报学报, 2006 (2): 301-308.

[9] Card S K, MacKinlay J D., Shneiderman B. Readings in information visualization: Using vision to think[G]. San Francisco: Morgan Kaufmann Publishers, 1999: 285-286.

[10] 邱均平. 信息计量学[M]. 武汉: 武汉大学出版社, 2007.

[11] Kessler M. Bibliographic coupling between scientific papers[J]. American Documentation, 1963 (14): 123-131.

[12] Jarneving B. Bibliographic coupling and its application to research-front and other core documents[J]. Journal of Informetrics, 2007 (1): 287-307.

[13] 邱均平. 信息计量学(六)第六讲 文献信息作者分布规律——洛特卡定律[J]. 情报理论与实践, 2000 (6): 475-478.

靖培栋 北京师范大学管理学院教授、博士生导师,博士。通讯地址:北京市。邮编:100875。

邱 威 牟晟斓 北京师范大学 2007 级情报学硕士研究生。通讯地址同上。

(收稿日期:2010-03-07;修回日期:2010-05-04)