DOI:10.16353/j.cnki.1000-7490.2009.09.019

● 苏 娜[1,2,3]，张志强[1,2]

（1.中国科学院 国家科学图书馆兰州分馆，甘肃 兰州 730000；2.中国科学院 资源环境科学信息中心，甘肃 兰州 730000；3.中国科学院 研究生院，北京 100049）

社会网络分析在学科研究趋势分析中的实证研究

——以数字图书馆领域为例

摘 要：本文验证了W.H.Lee提出的社会网络分析指标在学科研究趋势预测中应用的结论，利用社会网络分析的中心性指标对数字图书馆领域的研究趋势进行了分析和预测，认为“多媒体”可能成为数字图书馆领域未来的研究趋势之一。研究结果表明，利用中心性指标可以揭示出学科研究趋势，但是就目前的研究结果来看，单纯的社会网络分析指标不能很好地识别学科领域的研究趋势，应与其他方法相结合来进行趋势预测研究。

关键词：数字图书馆；社会网络分析；研究趋势

Abstract: This paper verifies the conclusion that Social Network Analysis (SNA) indicators put forward by W.H. Lee can be used to predict the research trend of the subject. The paper analyzes and predicts the research trend of the digital library field by the use of the centrality indicators of SNA, and believes that "multimedia" may become one of the research trends in future in the digital library field. The research results show that the centrality indicators of SNA can be used to reveal the research trend of the subject; however, in accordance with the present research results, pure SNA indicators can not identify the research trend of the subject satisfactorily, they must be combined with the other method in predicting the research trend of the subject.

Keywords: digital library; social network analysis; research trend

社会网络分析（Social Network Analysis，SNA）是社会学中用于研究社会成员之间关系的一种定量研究方法，距今已有70多年的历史。至今，社会网络分析法已经形成了一系列专有的概念和指标，并被广泛应用于社会学以外的其他众多学科。

作为一种研究人与人之间社会关系的定量化方法，社会网络分析法最初在情报学中的应用大多专注于研究社会人之间的关系，主要应用于合著网络分析、竞争情报中的人际网络分析、知识管理中如何利用社会网络理论促进成员之间的知识共享[1]。该方法还被扩展到引文网络分析当中，以作者之间的引用关系或同被引关系为网络连线，判断作者之间的相关关系。

近几年，一些学者开始尝试将社会网络分析的方法和指标扩展到人际关系之外的网络分析中，用来识别和分析某一学科领域的研究趋势。刘则渊等对国际科学学与科学计量学领域中6种核心期刊的高频关键词的共词网络进行分析，利用社会网络分析的方法对共词网络进行聚类，指出科学学与科学计量学的研究主题，并在此基础上对科学学研究的若干热点、未来的研究方向进行了预测[2]。陈超美利用社会网络分析中的中间中心性指标（Betweenness）识别共引网络中的关键节点，结合突发发现算法来探测领域未来研究趋势，并对Mass Extinction和Terrorism两个领域进行了实证研究[3]。W.H.Lee构建了信息安全领域的关键词共现网络，利用社会网络分析中的中心性指标（Centrality）发现当前研究热点，并对未来可能的研究趋势进行了分析[4]。

W.H.Lee采用历史经验法对数据进行分析，应用当前研究热点在刚开始出现时的中心性特征来判断当前的研究中哪些可能成为未来的研究趋势。通过数据分析，该作者认为中心性指标中的度（Degree）、中间中心性和接近中心性（Closeness）3项指标不但可以用于发现某一学科领域中当前的研究热点，而且还能够用于识别未来的发展趋势。

学科领域中新兴发展趋势的特征从社会网络分析角度可以被描述为：节点的度和接近中心性两项指标的值较低，而中间中心性指标的值较高。在研究中，W.H.Lee

并没有把这些指标量化，而只是对这些指标的特点进行了定性描述。

本文的研究目的有两个，一是在数字图书馆研究领域验证这种思路（历史经验法与中心性指标相结合）的可行性，并分析各种中心性指标的特点；二是利用这种方法尝试对数字图书馆领域的研究趋势进行分析。

1 社会网络分析的中心性指标

社会网络分析方法中涉及不同的概念和指标，在情报学领域应用的指标主要为中心性指标，用来衡量节点在网络中处于中心位置的程度。在网络中处于核心地位的点，会具有很高的中心度。在 L. C. Freeman的著述中有一整套对中心度进行测量的指标[5]，包括度、中间中心性和接近中心性。

度。又称“关联度”或“局部中心度”（Local Centrality），是指与某节点直接相连的节点个数。在社会网络分析中，一般称由一条线连接的节点是相互“邻接的”（Adjacent），与某个特定节点相邻的那些点称为该点的邻域（Neighborhood），邻域中的总点数称为度数。一个点的度就是对其“邻域”规模大小的一种数值测度[5]。在有向图中，某节点的度包括两个不同方面，分别称为点入度（Indegree）和点出度（Outdegree），一个点的点入度指的是直接指向该点的点数总和，点出度指该点所直接指向的其他点的总数。

中间中心性。用来测量的是一个点在多大程度上位于图中其他点的“中间”，即节点位于其他节点间最短路径的次数。一个度相对比较低的节点可能起到重要的中介作用，因而处于网络的中心。一个点的中间中心性测量的是该点对应的行动者在多大程度上成为“掮客”或者“中间人”，能在多大程度上控制他人。

接近中心性。又称“整体中心度”（Global Centrality），用来测量不同点之间的“距离”。两个点是由一条包含不同线的路径（Path）连在一起的，路径的长度用组成该路径的线数来测量。在一个网络图中，任何两点之间的最短距离称为“捷径”。接近中心性就是指节点与其他各个节点之间的捷径距离之和。如果一个点与其他许多点的距离都很短，则称该点是整体中心点，该点的接近中心性最低。

研究趋势是指随着时间逐渐引起人们兴趣，并被越来越多的学者讨论和应用的主题领域[6]。作为未来研究趋势的主题领域在开始出现时会表现出一些特征。从中心性分析的角度看，它所具有的特征是：

1）度的值比较低。度的大小代表与某节点相连的其他点的个数，度越大，表明与该节点关联的其他节点越多。从理论上来讲，新兴研究趋势在开始出现的时候，与其他主题节点的关联相对较少，在整个网络中处于边缘的位置，因此它的度即局部中心性较低，与其他节点连接线的数量较少。

2）接近中心性的值比较低。当一个节点的度比较低的时候，与该点连接的线比较少，相应的计算出来的节点的接近中心性也会比较低。接近中心性和度成正相关的关系。

3）中间中心性的值较高。一个新兴的潜在的研究热点在开始出现时，大多是位于其他主题领域的边缘位置，而且它往往起到了连接不同领域的作用，所以未来研究趋势的中间中心性是比较高的。

W. H. Lee通过实验证明了这一点，但是他并未对指标值的高低进行量化，只是通过实证方法推导出结论。此外，W. H. Lee仅在信息安全领域进行了实证分析，该方法的适用性需要进一步的验证。

2 数据与方法

为了验证社会网络分析指标在研究趋势分析中的应用，进一步验证 W. H. Lee所得出结论的可靠性和适用性，本文以数字图书馆为例，采用 W. H. Lee的实验思路进行检验，并对研究结果进行分析。

2.1 数据来源

研究以 Web of Science数据库为数据来源，以“Digital Library”为主题词进行检索，为了减少该数据库中综述、述评、会议文摘等类型数据对所构建数据集的影响，故将文献类型限定为“Article”和“Proceedings Paper”，共得到 2 977篇文献，去重后精简为 2 968篇。在构建的数据集中，作者给定的关键词共 4 075个，经过清洗去重，得到 3 873个。为了有效地对研究领域的结构进行分析，选取出现次数为 5次以上的 118个关键词构建共现矩阵。

2.2 方法及实施过程

网络可以分为随机网络（Random Network）和无标度网络（Scale-free Network）两种。传统的随机网络连接是随机设置的，大部分节点的连接数目会大致相同，即节点的分布方式遵循钟形的泊松分布，有一个特征性的“平均数”。

连接数目比平均数高许多或低许多的节点都极少，随着连接数的增大，其概率呈指数式迅速递减，故随机网络亦称指数网络。无标度网络的概念由 Barabasi提出。在无标度网络中，存在拥有大量连接的集散节点，节点与节点之间的连接分布遵循幂指数分布定律，其中大部分的节点只与少数节点连接，而少数节点则拥有大量的连接[7]。

在进行指标分析之前需要验证数据的特性，这是指验证数据是否符合幂指数分布，从而判断共词网络是否符合无标度网络的特征。只有当共词网络成幂指数分布时，整个网络中才会存在少数的节点连接着整个网络，才能从中找出具有重要连接作用的节点。在共词网络中，这些节点就可以视为研究热点。

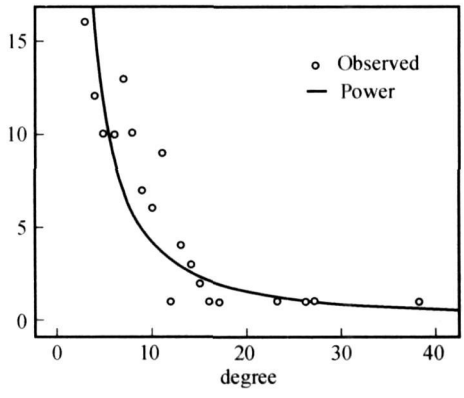

图 1 数字图书馆共词网络幂指数分布检验

从图 1可以看出，构建的数据集基本上是符合幂指数分布的，即关键词的共现网络是一个无标度网络（$Y=116.130 \cdot X^{1.441}$）。

因为网络节点的度数分布是符合幂指数分布的，因此可以进一步分析该网络的各项中心性指标。为了筛选出数字图书馆领域中的几个研究主题，笔者对共词网络进行了聚类分析。

由于数字图书馆领域各研究主题之间的联系相对比较密切，通过层级聚类法和 k值聚类都不能很好地将这些关键词进行聚类，因此本文选择在社会网络分析软件 Ucinet中用最优方法进行聚类，共指定为 20个类，聚类结果显示第 4类为天文学的相关主题类，是数据分析中的噪点，故将其删除。聚类之后需要为每一类赋予类名。类名赋予的方法有两种：一是以每类中出现次数最多的词作为类名；二是根据每一类所包含的内容为该类别赋予一个新词来概括其主要内容。由于第一种类名赋予方法比较简便，易于操作，所以本文同样也采取了这种方法。得到的 19个新大类分别为：archive，web service，OAI-PMH，information visualization，information retrieval，collaboration，indexing，architecture，digital signal processing，metadata，multimedia，image processing，browsing，digital preservation，education，evaluation，visualization，knowledge management，libraries。19个新大类代表了数字图书馆领域的 19个研究主题（关键词类）。将这些主题进行二次共现，如图 2所示。

在图 2的数字图书馆研究主题的关键词类共现网络中，19个类的中心性指标如表 1所示。

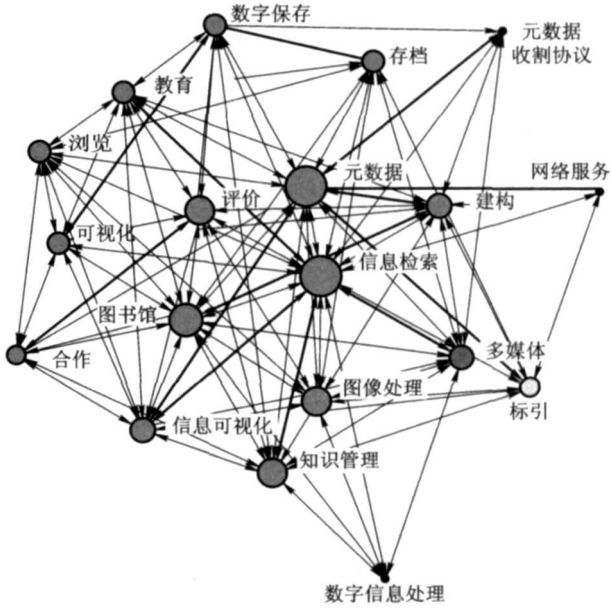

图 2 数字图书馆研究主题的关键词类共现网络

表 1 数字图书馆领域 19类主题领域的中心性指标

研究主题	度	中间中心性	接近中心性
存档（Archive）	10	4.124	48
网络服务（Web Service）	5	0.143	53
元数据收割协议（OAI-PMH）	5	0.268	53
信息可视化（Information Visualization）	12	2.624	46
信息检索（Information Retrieval）	18	16.105	40
合作（Collaboration）	9	0.49	49
标引（Indexing）	9	3.143	49
建构（Architecture）	12	4.878	46
数字信号处理（Digital Signal Processing）	5	0	53
元数据（Metadata）	18	18.055	40
多媒体（Multimedia）	11	3.619	47
图像处理（Image Processing）	13	4.498	45
浏览（Browsing）	10	1.079	48
数字保存（Digital Preservation）	10	2.584	48
教育（Education）	11	1.649	47
评价（Evaluation）	13	4.127	45
可视化（Visualization）	11	1.257	47
知识管理（Knowledge Management）	13	8.073	45
图书馆（Libraries）	15	6.16	43

通过对网络和数据进行分析，并与相关专家商榷，笔者认为在数字图书馆领域的研究中，元数据是比较具有代表性和影响的主题，是目前的研究热点，可以作为中心节

点 (Hub)。从表 1中 19大类主题的中心性指标可以看出, Metadata的度非常高, 在 19类中与 Information Retrieval同为最高值; 但是它的接近中心性的值却是最低; 中间中心性的值在 19大类中也是最高的。这与 W. H. Lee选择的中心节点的中心性值在整个类别中的位置大致是吻合的。

通过对数据的分析得知, 在本文的数据集中, 元数据最早出现在 1996年, 所以选取 1996— 2000年共 5年的数据, 按照上述方法重新构建共词网络, 最后得出类的共现图, 如图 3所示 (126个关键词聚成 20个类)。

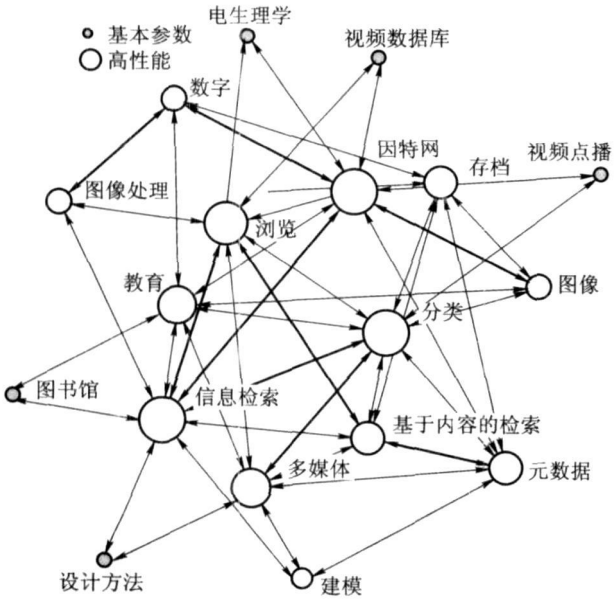

图 3 1996— 2000年数字图书馆研究主题的共现图

在 1996— 2000年的数字图书馆主题的共词网络中, 各关键词类, 即研究主题的中心性指标如表 2所示。

将 20个主题的中心性指标的每一项指标都依值的大小分为 3个等级, 分别为“高”、“中、”“低” 3档, 通过对中心性指标值进行分析可以看出, 元数据在刚开始出现的几年中, 度的值在第二等级, 处在中间偏上的位置, 而它的接近中心性的值虽然也位于第二等级, 但处在中间偏下的位置, 而中间中心性的值处于中间位置, 其值相对偏低一些。

这样的结果与前面的理论分析以及 W. H. Lee得出的结论是存在一定差别的。

接近中心性的值偏低, 与前面的结论基本是一致的; 但是中间中心性的值却并不是像理论上预测的比较高, 而是偏低; 度也并不是像预测的那样比较低, 而是处于中间偏上的位置。

根据上述中心性指标特点, 在图 2的关键词类共现网络中也找到具有类似特点的节点, 即节点多媒体。据此认为如果利用共词网络对当前数字图书馆研究中未来的趋势进行预测, 则多媒体可能成为未来的研究热点之一。但多媒体是一个多概念的集合名词, 其中包含了一系列与多媒体有关的研究。

表 2 1996— 2000年数字图书馆研究主题的中心性指标

研究主题	度	中间中心性	接近中心性
多媒体 (Multimedia)	7	9. 743	87
视频数据库 (Video Database)	2	0	96
数字 (Digital)	4	2. 167	92
元数据 (Metadata)	6	5. 775	89
因特网 (Internet)	9	25. 969	85
图像 (Images)	4	1. 333	92
信息检索 (Information Retrieval)	9	25. 613	85
视频点播 (Video on Demand)	2	0. 167	96
图书馆 (Libraries)	2	0	97
基于内容的检索) (Content-based Retrieval)	6	2. 643	88
基本参数 (Fundamental Parameters)	0	0	400
建模 (Modeling)	3	0. 56	96
存档 (Archive)	6	6. 292	92
教育 (Education)	7	11. 245	87
设计方法 (Design Methodology)	2	0. 167	98
高性能 (High-performance)	0	0	400
图像处理 (Image Processing)	4	2. 583	91
分类 (Classification)	9	14. 911	85
电生理学 (Electrophysiology)	2	0	96
浏览 (Browsing)	8	16. 833	86

2.3 分析

通过上述分析, 笔者认为 W. H. Lee的研究思路和方法是可以应用到学科发展趋势预测中的, 根据 W. H. Lee的方法, 多媒体可能成为未来的研究热点。但由于研究的局限, 这里再对上述两个结论予以进一步说明。

1) 关于数字图书馆研究趋势的结论。基于指标特征得出的多媒体可能成为未来研究热点之一的结论与目前研究内容的发展是一致的。M. Y. Tsay对数字图书馆研究的引证期刊和被引期刊的主题进行了比较, 结果表明, 引证期刊的主题主要集中在计算机与信息技术的应用。而大部分被引期刊关注比较多的主题, 如图书馆自动化、信息检索、信息服务、主题标引等没有出现在引证期刊中[8]。这说明数字图书馆研究正在由理论向实践转化, 对数字图书馆的研究已从初始的概念研究逐步转向具体的技术实现。在未来的若干年内, 多媒体将对图书馆产生深刻的影响[9]。

它不仅将成为信息资源的主要形式, 而且也将成为图书馆服务的基础。随着文本、图像、视频处理技术的相对

成熟，多媒体技术将成为数字图书馆领域关注的重点内容之一，可能成为未来数字图书馆领域的研究趋势。但不可否认的是，研究趋势的预测是一项具有前瞻性的工作，结论需要相关领域专家的进一步讨论，也需要时间证明。

2）在验证“元数据”兴起的最初几年的中心性指标时，与前面的理论分析和 W. H. Lee得出的结论存在一定差别。利用中心性指标进行数字图书馆领域研究趋势分析的结果显示，接近中心性的值偏低，与前面的结论基本是一致的；但是中间中心性的值却并不是像理论上预测的比较高，而是偏低；度中心性也并不是像预测的那样比较低，而是处于中间偏上的位置。出现这种情况的原因可能有 3个。

一是在实验过程中对数据的处理方法不完全一致。W. H. Lee对共词网络进行聚类采用了层次聚类法，而在本文的数据操作过程中，由于数字图书馆领域各子主题之间联系比较紧密，采用层次聚类法没有聚出理想的类，所以采用了最优聚类法。此外，对共现矩阵的处理方式也不完全相同。W. H. Lee对关键词共现矩阵的处理采用了较为普遍、通用的方法，进行了皮尔森相关处理；而 Leydesdorff则认为如果是对称矩阵，则没有必要进行相关性转换，直接使用原始矩阵进行分析[10]。笔者也认同这种观点。因为本文的关键词类共现矩阵为对称矩阵，所以在将其导入 Ucine进行聚类、可视化分析之前，没有对矩阵进行相关性转换处理，而是利用初始矩阵。

二是学科领域的差异。W. H. Lee选择信息安全领域为研究对象进行分析，而本文选择了数字图书馆领域为研究对象，两个不同学科领域在发展速度和特点上会有很大的不同，可能造成了实际的分析预测过程中指标值的不一致性。

三是中心性指标的限制。在学科研究趋势预测的研究中，中心性指标并不是一个确定的量化的值，而是高、中、低等相对概念，这很难用一个具体的量化的阈值去限定，也可能造成了本文的研究结果与 W. H. Lee研究结果的差异。

在进行数据处理过程中，采用初始矩阵会对分析结果造成一定的影响，但因为相关性指标问题所带来的这种影响在实际研究中是比较小的[11]，所以在数据处理过程中的不一致理论上并不造成结果的差异。而学科领域的差异和对指标高或低的不同认识则可能是造成这种结论不同的原因。

3 结束语

本文验证了 W. H. Lee提出的社会网络分析指标在研究趋势预测中应用的结论，利用社会网络分析的中心性指标对数字图书馆领域的研究趋势进行了分析和预测，认为“多媒体”可能成为数字图书馆领域未来的研究趋势之一。

研究结果表明，利用中心性指标与历史经验法相结合的方法可以揭示出研究趋势，但是就目前的研究结果来看，单纯的社会网络分析指标不能很好地识别某领域的研究趋势，应与其他方法相结合来进行研究趋势预测。此外，在社会网络指标应用的过程中，要注意具体指标特点与具体领域相结合。那么，社会网络分析方法如何与其他方法相结合来识别学科领域的研究趋势，将是需要进一步研究的问题。□

参考文献

[1] 朱庆华，李亮．社会网络分析法及其在情报学中的应用[J]．情报理论与实践，2008（2）：179-183，174.

[2] 刘则渊，尹丽春．国际科学学主题共词网络的可视化研究[J]．情报学报，2006（5）：634-639.

[3] CHEN Chaomei. 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.

[4] LEE W H. How to identify emerging research fields using scientometrics: an example in the field of information security[J]. Scientometrics, 2008, 76(3): 503-525.

[5] 约翰·斯科特．社会网络分析法[M]．刘军，译．重庆：重庆大学出版社，2007.

[6] KONTOSTATHIS A, et al. A survey of emerging trend detection in textual data mining[EB/OL]. [2009-01-04]. http://dimacs.rutgers.edu/~billp/pubs/ETDArticle.pdf

[7] 车宏安，顾基发．无标度网络及其系统科学意义[J]．系统工程理论与实践，2004（4）：11-16.

[8] TSAY M Y. Subject change between citing and cited literature on digital libraries[J]. The Electronic Library, 2008, 26(5): 702-715.

[9] MITCHELL G A. Distinctive expertise: multimedia, the library and the term paper of the future[J]. Information Technology and Libraries, 2005, 24(1): 32-36.

[10] LEYDESDORFF L, VAUGHAN L. Co-occurrence matrices and their applications in information science: extending ACA to the Web environment[J]. Journal of the American Society for Information Science and Technology, 2006, 57(12): 1616-1628.

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

作者简介：苏娜，女，1983年生，博士生。发表论文6篇。

张志强，男，1964年生，研究员。发表论文150余篇，出版专著和编著7部，译著5部。

收稿日期：2009－04－13