网络信息计量学研究主题分析
——基于共词可视化方法

黄 莉[1] 李 江[2]
(1. 南京大学商学院,江苏南京 210093;2. 南京大学信息管理系,江苏南京 210093)

摘 要:本文从SCIE中挑选出网络信息计量学领域的20个热点关键词构建共词矩阵,借助共词可视化方法绘制了以"Webometrics"为例的Pajek网络图,并依据可视化图中节点的大小与连线的粗细分析了网络信息计量学的六大研究主题,分别为学科范畴、理论基础、研究对象、计量指标、研究工具、应用范围,并根据可视化图判断出理论基础、研究对象等。属于热点研究主题。

关键词:网络信息计量学;研究主题;共词可视化;Pajek网络图
中图分类号: G353.1 **文献标识码**: A DOI: 10.3772/j.issn.1674－1544.2008.04.002

Analysis of the Research Topics of Webometrics
——Based on Co－word Visualization

Huang Li[1], Li Jiang[2]
(1. School of Business, Nanjing University, Nanjing 210093;
2. Department of Information Management, Nanjing University, Nanjing 210093)

Abstract: This paper picks out 20 hot key words in the area of Webometrics to construct the co－word matrix, and protracts the Pajek networks graph by taking the methodology of Co－word Visualization, and based on the size of point and the thickness of linking line in the graph of visualization finally analyzes the six research topics of Webometrics, including discipline category, basic theories, research objects, metrics indicators, research tools, application and so on, in which basic theories and research objects are key topics.

Keywords: Webometrics, research topics, Co－word Visualization, Pajek networks graph

近十余年来,国内外网络信息计量学相关论文的数量分布如表1所示。SCIE中的相关论文代表了全世界该领域的优秀学术论文,而CNKI中的相关论文则代表了国内该领域几乎全部的学术论文。总体而言,国内外网络信息计量学的相关论文呈增长趋势,也印证了这一新兴学科快速发展的特征。笔者尝试以SCIE中的相关论文为依据,借助共词可视化方法,分析网络信息计量

表1 SCIE与CNKI中网络信息计量学相关论文在时间轴上的数量分布

年份	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	累计(篇)
SCIE	1	2	1	3	3	5	10	14	10	7	6	62
CNKI	0	0	1	2	7	14	13	12	23	37	25	134

第一作者简介:黄莉(1983－),女,南京大学2006级企业管理硕士研究生,研究方向是数据挖掘与人力资源管理。

收稿日期:2008年4月15日。

学的研究主题。

1 共词可视化方法

“共词可视化”是指将可视化技术与共词原理以及其他的定量方法相结合，通过分析词在文献中共同出现的现象，生成具有各种属性的可视化结果，如各种图形、图表或科学地图，以揭示知识领域结构、映射知识领域发展趋势[1]。

工具方面，目前最常用的是SPSS与Pajek。SPSS是三大统计分析软件之一，应用非常广泛，通过其聚类分析、因子分析、多维尺度分析、对应分析等功能可将共词分析的结果可视化。Pajek(在斯洛文尼亚语中是“蜘蛛”的意思)是一种基于Windows的将大型网络可视化的社会网络分析软件，通过可视化的展现，扩展了人类的视觉功能，允许人们对大量抽象的数据进行分析[2](可从http://vlado.fmf.uni-lj.si/pub/networks/pajek/免费获取)。

应用方面，周静怡等[1]根据该领域中出现的高频关键词之间的共词关系，通过3种方法——战略坐标、网络图和自相关地图生成相应的3种可视化结果，各有侧重地揭示了特定学科领域的研究热点和发展趋势。刘则渊等[3]采用可视化手段对国际科学学与科学计量学领域中6种期刊的高频关键词的共词网络进行了分析，利用社会网络分析的方法对共词网络进行聚类，研究表明，科学、技术和创新活动的研究是科学学与科学计量学的主题。

2 共词可视化过程

本文中的共词可视化过程采用Pajek作为工具，绘制以节点大小表示关键词的词频、以连线粗细表示关键词之间相关性的网络图。

2.1 数据获取

以“Webmetrics”、“Webometrics”(根据不同的构词法，该术语的拼写方式不同)为关键词在SCIE中搜索相关论文，然后提取各篇论文中的关键词，将各关键词的词频按降序排列，剔除其中专指程度低的关键词，如“science”等，剩余20个专指程度较高的关键词如表2所示，这20个关键词可以作为分析网络信息计量学研究主题的代表性关键词。各关键词的词频将在Pajek网络图中显示为结点的大小(表2)。

将这20个关键词作为共词矩阵的单元，两个关键词同时出现在一篇论文中，则视为这两个关键词共现一次，按此原理统计这20个关键词两两之间的共词次数，并构建共词矩阵。再将共词矩阵转化为相关矩阵，各关键词之间的相关系数将在Pajek网络图中显示为节点之间连线的粗细(图1)。

2.2 Pajek网络图绘制

若将20个关键词的相关数据全部导入Pajek，将构成错综复杂的网络图，难以辨认其中的

表2 网络信息计量学领域热点关键词

关键词	词频	编号	关键词	词频	编号
Webometrics	24	K01	Scholarly Communication	7	K11
World Wide Web	23	K02	Information Science	6	K12
Information	19	K03	Site Interlinking	6	K13
Web Sites	17	K04	Co-authorship	4	K14
Impact Factors	16	K05	Citation	3	K15
Bibliometrics	13	K06	Co-citation	3	K16
Search Engines	13	K07	Collaboration	3	K17
Web Impact Factors	13	K08	Crawler	3	K18
Citation Analysis	10	K09	Research Assessment Exercise	3	K19
Links	10	K10	University Web Sites	3	K20

	K01	K02	K03	K04	K05	K06	K07	K08	K09	K10	K11	K12	K13	K14	K15	K16	K17	K18	K19	K20
K01	24	8	7	10	8	8	9	3	6	4	4	0	2	1	0	1	0	2	1	1
K02	8	23	9	7	6	9	4	3	5	5	5	4	4	2	2	2	2	0	0	2
K03	7	9	19	4	5	4	5	5	4	6	2	0	2	1	1	2	0	3	1	1
K04	10	7	4	17	5	5	8	1	6	4	3	0	3	1	0	1	1	0	2	1
K05	8	6	5	5	16	4	5	1	5	4	0	0	1	0	0	0	0	0	1	1
K06	8	9	4	5	4	13	1	1	3	4	2	0	1	0	0	1	1	0	0	0
K07	9	4	5	8	5	1	13	2	4	3	3	1	2	1	0	0	0	1	2	1
K08	3	3	5	1	1	1	2	13	2	1	1	1	1	0	1	0	1	1	1	1
K09	6	5	4	6	5	3	4	2	10	3	2	0	2	0	0	0	1	1	1	0
K10	4	5	6	4	4	4	3	1	3	10	1	1	1	1	1	1	1	1	0	1
K11	4	5	2	3	0	2	3	1	2	1	7	2	3	0	0	1	0	0	1	0
K12	0	4	0	0	0	0	1	1	0	1	2	6	1	1	1	0	1	0	0	1
K13	2	4	2	3	1	1	2	1	2	1	3	1	6	0	0	0	0	1	0	1
K14	1	2	1	1	0	0	1	0	0	1	0	1	0	4	1	0	1	1	0	1
K15	0	2	1	0	0	0	0	1	0	1	0	1	0	1	3	0	1	0	1	1
K16	1	2	2	1	0	1	0	0	0	1	1	0	0	0	0	3	0	0	0	0
K17	0	2	0	1	0	1	0	1	1	1	0	1	0	1	1	0	3	0	0	1
K18	2	0	3	0	0	0	1	1	1	1	0	0	1	1	0	0	0	3	0	1
K19	1	0	1	2	1	0	2	1	1	0	1	0	0	0	1	0	0	0	3	0
K20	1	2	1	1	1	0	1	1	0	1	0	1	1	1	1	0	1	1	0	3

图 1　代表性关键词的共词矩阵

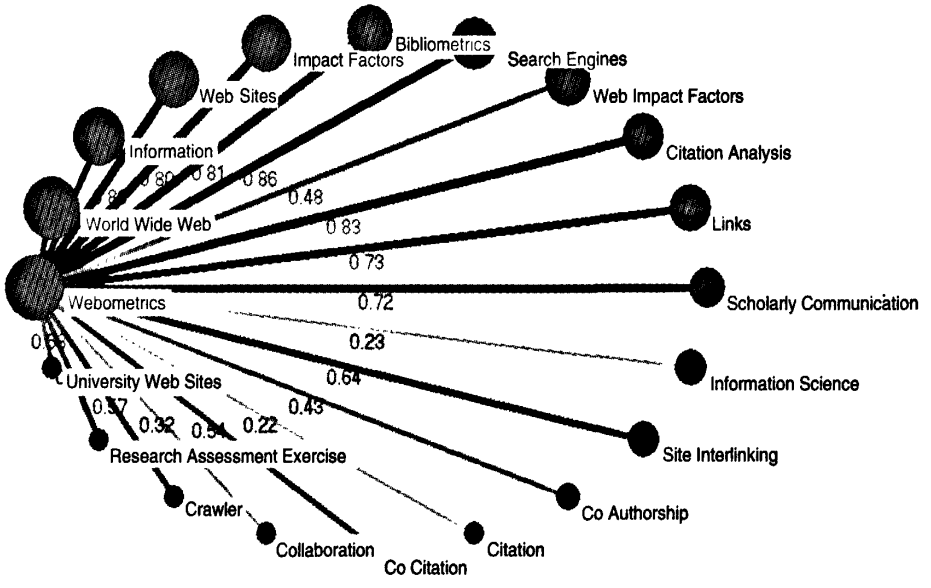

图 2　20 个网络信息计量学领域关键词的共词可视化图

关系。笔者尝试以“Webometrics”为例，用 Pajek 绘制它与其他 19 个关键词之间的共词关系网络图。将上一节中获取的“Webometrics”相关数据按照 Pajek 所需的格式导入 Pajek 中，运行后得到图 2 所示的结果。

图 2 中显示的网络图由节点与连线构成，从外形上看，这 20 个关键词节点是以“Webometrics”为中心，其余 19 个关键词节点从其辐射而出。各节点旁边均有标签标明其含义，节点的大小代表该关键词词频的高低，连线的粗细代表关键词关联的强弱，具体数值也标在各连线旁。

3　网络信息计量学研究主题分析

可视化图可以直观地揭示各关键词在网络信息计量学领域的重要位置以及它们之间的潜

在关联。借助图2,笔者将网络信息计量学的研究主题分为学科范畴、理论基础、研究对象、计量指标、研究工具和应用范围6个部分。

3.1 研究主题分析

3.1.1 学科范畴

代表性关键词——Information(信息)、Information Science(信息科学)。网络信息计量学从其研究内容——网络信息的角度看,隶属于信息科学,是信息科学范畴内自1997年才兴起的一门新兴学科。

"Information Science"节点位于可视化图的下半区,与"Webometrics"之间的连线较细。这表明学科范畴这一主题并非核心,因为自网络信息计量学诞生以来,学科特征明显,其学科范畴未有争议。

3.1.2 理论基础

代表性关键词——Impact Factors(影响因子)、Bibliometrics(文献计量学)、Citation Analysis(引文分析)、Co-authorship(合著)、Citation(引文)、Co-citation(共引)。文献计量学是网络信息计量学的理论基础。具体而言,网络信息计量学中的链接分析来源于文献计量学中的引文分析;网络信息计量学中的网络影响因子来源于文献计量学中的期刊影响因子;网络信息计量学中的共链分析(Co-link Analysis)来源于文献计量学中的共引分析(Co-citation Analysis)。

这一主题的关键词节点之多,"Impact Factors"、"Bibliometrics"、"Citation Analysis"等节点与"Webometrics"之间连线之粗是其他主题无法比拟的,这也表明了理论基础这一主题在网络信息计量学研究中的地位。事实的确如此,自网络信息计量学诞生至今,文献计量学理论为其提供了理论基础,沿用文献计量学理论的合理性也一直是研究的焦点。

3.1.3 研究对象

代表性关键词——World Wide Web(万维网)、Web Sites(网站)、Links(链接)、Site Interlinking(站间链接)、University Web Sites(大学网站)。吴华香提出"网络信息计量学是互联网上的文献计量学"[4]。如今看来,这一观点并非准确,但依然描述出了网络信息计量学的一些特征,如延用了文献计量学的理论方法,但将研究对象由文献环境中的文献、引文等换成了网络环境中的链接、网站等。具体而言,网络信息计量学初期的研究对象主要为学术网络信息,如大学网站、大学网站中的链接以及站间链接等。

这一主题的关键词的节点均较大、与"Webometrics"之间连线均较粗,这也旗帜鲜明地揭示了网络信息计量学的研究对象,不容混淆。

3.1.4 计量指标

代表性关键词——Web Impact Factors(网络影响因子)。网络影响因子这一指标来源于文献计量学中的期刊影响因子[5],承袭"链接代表认可"这一观点,主要用于评价网站的影响力。自1998年P. Ingwersen提出这一指标后,该指标频繁出现在链接分析的理论与应用研究中,因此,该节点出现于可视化图的上半区,但因为该指标适用性差、算法上存在缺陷、假设前提难以成立[6],所以未被应用于网站评价实践中。在可视化图中,该节点与"Webometrics"之间连线较细可以说明这一点。

3.1.5 研究工具

代表性关键词——Search Engines(搜索引擎)、Crawler(爬虫)。初期的网络信息计量学研究主要依赖于商业搜索引擎,如Altavista、Alltheweb、Google等。商业搜索引擎的作用在于通过特定的指令获取网站的链接数据,包括入链数(Inlinks)、出链接(Outlinks)以及网站内部页面总数。后来,学者们逐渐发现商业搜索引擎用作网络信息计量工具的缺陷——不一致性。不一致性主要包括以下两种情况:①不同时刻(间隔几小时或几天),同一检索式的检索结果不一致;②(几乎)同一时刻,检索式"A and B"与"B and A"的检索结果不一致,或"A and B"与"A and not B"检索结果中记录总数之和与"A"检索得到的记录总数不一致。于是,学者们纷纷借助爬虫获取特定网站集中的链接数据,或者自行开发网络信息计量工具

获取链接数据，如 Lei Cui 开发的 Checkweb、段宇锋开发的 Webstat 等。

可视化图中，“Search Engines”与“Crawler”节点的大小及其与“Webometrics”之间连线的粗细，清晰地显示出了这两种工具在网络信息计量学领域应用的范围大小。

3.1.6 应用范围

代表性关键词——Scholarly Communication（学术交流）、Collaboration（合作）、Research Assessment Exercise（研究评估实践）。从关键词上看，网络信息计量学主要应用于科学评价与学术交流两个领域。在科学评价方面，网络信息计量学中的链接分析方法作为一种定量评价方法，在网络信息资源评价方面应用广泛，如入链数、网络影响因子、PageRank 算法等指标都能反映出网络信息资源的质量；在学术交流方面，邱均平等提出，通过对因特网上的有关各学科的站点、讨论组、电子期刊等的计量分析，可以掌握科学信息在网络上的分布；通过对相关网站之间的链接用于被引分析乃至利用专用软件分析特定对象的电子邮件使用情况，可以了解网上的科学信息交流情况[7]。

可视化图中，这一主题的 3 个节点均不在上半区，且各节点与“Webometrics”之间连线均较粗，这表明这一主题在网络信息计量学虽不算核心，但随着互联网的发展，网络信息计量学的应用研究将越来越受关注。

3.2 可视化效果分析

共词可视化方法的理论依据是：共现关键词之间存在某种关联，共现次数越多，这种关联越强，并将这些关键词及其关联强度用节点的大小和连线的粗细描绘成图。这种定量、可视的方法直观地揭示了网络信息计量学领域的六大研究主题，节点与连线中隐含着许多能反映该领域研究现状的信息。

3.2.1 节点大小反映各节点关键词的受关注程度

图 2 中，从 Webometrics 开始，沿顺时针方向，各节点从大到小依次排列。各节点的大小代表该关键词在样本中出现频率的高低。频率越高的词在该领域中受关注程度越高（前提是按照上文的方法剔除专指程度较低的关键词）。在网络信息计量学研究中，World Wide Web、Web Sites、Search Engine 等受关注程度较高，因为 World Wide Web 是网络信息的载体，Web Sites 是网络信息的组织形式，Search Engine 是获取网络信息的工具，在实证研究中，这些关键词几乎都必不可少，而 Citation、Crawler、University Web Sites 等则相对次要，因为 Citation 仅仅会在探讨网络信息计量学起源的理论研究中被提到，Crawler 是极少数学者对商业搜索引擎检索效果不满意时才用到的工具，University Web Sites 是一类特殊的研究对象——学术性网络信息，这些关键词只在少数论文中被提到。

由此看来，图 2 所示的可视化图直观地揭示了网络信息计量学领域各专指程度较高的关键词（20 个）的受关注程度——从 Webometrics 开始，沿顺时针方向依次降低。

3.2.2 连线粗细反映各关键词与 Webometrics 的关联强度

连线粗细是指两个关键词之间的共现强度，共现强度在理论上是指关联强度。因此，图 2 中连线越粗代表该关键词与 Webometrics 之间的关联越强，反之亦然。

事实上，图 2 所示的节点较大的关键词并不能反映网络信息计量学领域的研究热点，即受关注程度高并不必然是研究热点。例如，World Wide Web 的受关注程度很高，但很难说 World Wide Web 是网络信息计量学领域的研究热点，因为在任何与互联网紧密相关的学科中，World Wide Web 出现的频率都很高。这一点也是词频统计法（完全依照词频判断某领域研究热点、研究主题等）的缺陷。

笔者认为，共词强度可以弥补词频统计法的这一缺陷，可以辅助判断某领域的研究热点。按照这一观点，可以从图 2 中直观地判断出上述六大研究主题中的研究热点：理论基础、研究对象等与 Webometrics 之间的连线较粗，可认为是热

点研究主题。

4 结束语

本文借助共词可视化的方法分析了网络信息计量学领域的六大研究主题，其中，理论基础、研究对象等属于热点研究主题。数据来源方面，SCIE中的网络信息计量学相关论文可以代表世界范围内的重点研究成果，也可以反映出网络信息计量学领域的研究现状，在此基础上的共词可视化分析得出的该领域的研究主题可信度高；关键词选取方面，剔除泛指关键词后，剩余20个关键词的专指程度较高，能反映出网络信息计量学领域的研究热点，但数据偏少，反映出的热点并不全面，这也是本文的不足之处；可视化图实例方面，以“Webometrics”为中心绘制的可视化图直观地反映出了与“Webometrics”共现的其他关键词所处的位置，直观地揭示了网络信息计量学领域的研究现状。共词可视化方法可作为文献计量学研究方法，定量而客观，可直观地揭示各学科领域研究现状。如果能结合专家的主观分析，这种方法将更有说服力。

参考文献

[1]周静怡，孙坦，陈涛．共词可视化：以人类基因组领域为例[J]．情报学报，2007，26(4)：532－537.

[2]于琪．共现分析法研究[D]．武汉：武汉大学，2007.

[3]刘则渊，尹丽春．国际科学学主题共词网络的可视化研究[C]．首届中国科技政策与管理学术研讨会2005年论文集(上)，2005.

[4]吴华香．网络计量学：互联网上的文献计量学[J]．图书馆杂志，2001(1)：33－36.

[5] Peter Ingwersen. The calculation of web impact factors[J]．Journal of Documentation，1998(2)：236－243.

[6]李江．链接分析工具研究[D]．武汉：武汉大学，2007.

[7]邱均平，陈敬全．网络信息计量学及其应用研究[J]．情报理论与实践，2001(3)：161－163.