我国数字图书馆研究论文（2005-2009）的统计分析——社群分析*

□ 张鹏 王继民 王建冬 / 北京大学信息管理系 北京 100871

摘要：基于万方收录的数字图书馆领域研究论文，对2005—2009年数字图书馆研究领域的文献、研究机构、作者和关键词进行网络分析。重点分析了该领域文献的机构合作、作者合著、关键词共现情况，得到了三个网络的社群分布图，对核心机构和作者进行了分析，关键词分析显示，研究热点主要集中在元数据、本体、个性化服务、数字版权和资源建设等方面。

关键词：数字图书馆，社会网络分析，机构合作，作者合著，共词网络
DOI：10.3772/j.issn.1673—2286.2010.03.017

引言

社群，即社会群体，它是社会学的基本概念之一，由于研究的角度不同， 对其定义有一定差异。从结构功能的角度定义，它指的是由两个或两个以上的具有共同认同和团结感的人所组成的人的集合，群体内的成员相互作用和影响，共享着特定的目标和期望。从广义上说，它既包括规模较小、交往密切而关系松散的群体（如朋友圈、俱乐部），也包括规模较大并且高度组织化的群体（如企业、学校、政府机关）[1]。

社会网络分析提出了大量概念和方法来观测网络结构类型、鉴别相互关系、分析网络成员行为体现的社会结构。社会网络分析家巴里•韦尔曼指出：“网络分析探究的是深层结构——隐藏在复杂的社会系统表面之下的一定的网络模式”[2]。在图书情报领域中的应用主要集中在共词分析、合著分析、引文分析等[3-8]方面，本文将机构、科研人员、文献关键词分别作为不同的“群体”，利用社会网络分析软件Pajek，对数字图书馆领域2005年至2009年五年的文献数据进行机构合作、作者合著和关键词共现分析。

1 数据来源和分析指标介绍

1.1 数据来源

本文所使用的数据是由万方数据股份有限公司收录的，包含国内图书馆学、情报学、档案学、编辑出版专业、新闻传播专业、计算机技术专业和综合类（各种学报等）的206种核心期刊的所有文献数据，共440594条记录，文献发表年份为1978-2009年，但前期数据收录不全。每条记录由标识号、题名、作者、机构、关键词、摘要、刊名、发表年、基金项目等字段组成。从以上文献数据中检索题名字段和关键词字段中包含“数字图书馆”的记录，并且设定发表年为2005-2009年的文献，得到记录1997条，作为国内数字图书馆研究的核心论文。选取这些记录中的机构、作者和关键词字段，并对其进行相应的处理，转换为适合Pajek分析的数据格式。

1.2 分析指标

经过多年的发展，社会网络分析已经形成了比

* 本文受北京大学前瞻性研究课题“基于网络使用的中文科技期刊及论文评价机制和方法研究”资助。

较完整的理论体系，涉及的概念和术语很多。本文选取了一些指标对数字图书馆的三个网络进行分析和解释，对这些指标进行简单的介绍，如下所示[2, 9]。

（1）密度（density）：网络的密度描述的是一个图中各个节点之间联系的紧密程度，对于无向网络来说，计算公式为$2l/(N^2-N)$，其中，l为总的连线数，N为总的节点数。

（2）度（degree）与度分布：在网络图中，如果两个节点由一条线相连，则称这两个点为“相邻的”。与某点相邻的那些点称为该点的“邻点”，一个点的邻点的个数称为该点的度数。网络中节点的度的分布情况可用分布函数P(k)来描述，PP(k)表示的是网络中度数为k的顶点的个数占顶点总个数的比例。

（3）直径和平均路径长度：网络中两个顶点i，j之间的最短路径定义为所有连通（i，j）的通路中，所经过的其他顶点最少的一条或几条路径。两个顶点i，j之间的距离d_{ij}定义为i，j之间最短路径上的边数。网络的直径（diameter），定义为网络中任意两个顶点之间距离的最大值。网络的平均路径长度（average path length），定义为网络中任意两个顶点之间距离的平均值，它描述了网络中节点间的分离程度，即网络有多小。计算公式为：

$$L=\sum_{i>j} d_{ij} / C_N^2$$

其中，L代表平均路径长度，N代表节点总数。

（4）介数（betweenness）：一个顶点v的介数定义为网络中所有的最短路径之中，经过v的数量，它反映了顶点v在多大程度上控制其其他顶点之间的交往。

（5）成分（component）：网络的成分是指最大的连通子图，一个网络由一个或多个成分组成，成分内部的所有节点都是通过路径相连的，成分之间是没有路径连通的。

2 我国数字图书馆研究机构合作网络分析

现代科学从小科学时代进入大科学时代，科研合作的规模和范围不断扩大，其结果通常表现为研究者共同署名在学术期刊上发表论文，因此，学术期刊是了解科研合作交流情况的重要途径。科研合作分析的研究主要是构建跨机构、跨省区、跨国家等的科研合作网络，以此为基础分析机构和区域的科研合作状况。

2.1 机构合作网络整体分析

对近5年国内数字图书馆研究机构的合作数据进行处理，其中，研究机构作为节点，研究机构之间合作发表文章作为边，合作发表文章的数量作为边的权值，得到相应的机构合作网络。机构合作网路的节点总数为765个，边数为556个，密度为0.000951，可见，各个研究机构在数字图书馆领域中合作很少。

对机构合作网络中各个节点的度数和介数进行统计分析，得到度数和介数最高的前10所研究机构，如表1所示。根据度数的排名，与其他机构合作最多的是中国科学院文献情报中心，紧接着是中国科学技术信息研究所、北京大学、吉林大学和武汉大学等机构。根据介数的排名，在机构合作网络中处于中心位置的是中国科学院文献情报中心，然后是北京大学、中国科学技术信息研究所、武汉大学、中国科学院成都文献情报中心和四川大学等。其中，吉林大学的度数排名是第4，而介数排名却不在前10名，原因是它处于分支网络的中心，与主干网络不连通。

对机构合作网络进行成分划分，并根据每个成分所包含的节点数进行分类统计，如表2所示，共得到540个成分，最大的成分只有105个节点，并且有472个成分只有一个节点，即整个网络中有472个孤立点，由此可见，整个机构合作网络非常松散。

2.2 机构合作网络主要成分分析

由于机构合作网络非常松散，我们仅选取包含节点数最多的两个成分做进一步的分析，它们的节点数量分别为105、17。如图1所示，黄色节点及其连线所构成网络是机构合作网络中的最大成分，即成分1，该网络的直径为11，平均路径长度为4.84。从图中可以明显看出，与其他机构合作较多的是中国科学院文献情报中心、中国科学技术信息研究所、北京大学和武汉大学，其中，处于整个网络中心位置的是中国科学院文献情报中心和北京大学，这与介数排名所得到的结果一致，而国家图书馆、南京大学和四川大学等机构处于局部网络的中心位置，介数排名也都在前10名，但相对靠后。从整个网络来看，机构间的合作关系显现出很强的地域性特征，即各个地区范围内的合作较多。

绿色节点及其连线所构成的网络是机构合作网络

表1 度数和介数最高的前10所研究机构

排名	度数	研究机构	介数	研究机构
1	16	中国科学院文献情报中心	0.01164	中国科学院文献情报中心
2	15	中国科学技术信息研究所	0.00914	北京大学
3	13	北京大学	0.00736	中国科学技术信息研究所
4	13	吉林大学	0.00577	武汉大学
5	12	武汉大学	0.00507	中国科学院成都文献情报中心
6	8	南京大学	0.00505	四川大学
7	8	国家图书馆	0.00328	华东师范大学
8	8	苏州大学图书馆	0.00300	清华大学图书馆
9	7	四川大学	0.00240	南京大学
10	6	清华大学图书馆	0.00233	国家图书馆

表2 成分分类统计

节点数	1	2	3	4	5	10	11	17	105	总计数
成分频次	472	50	8	4	2	1	1	1	1	540

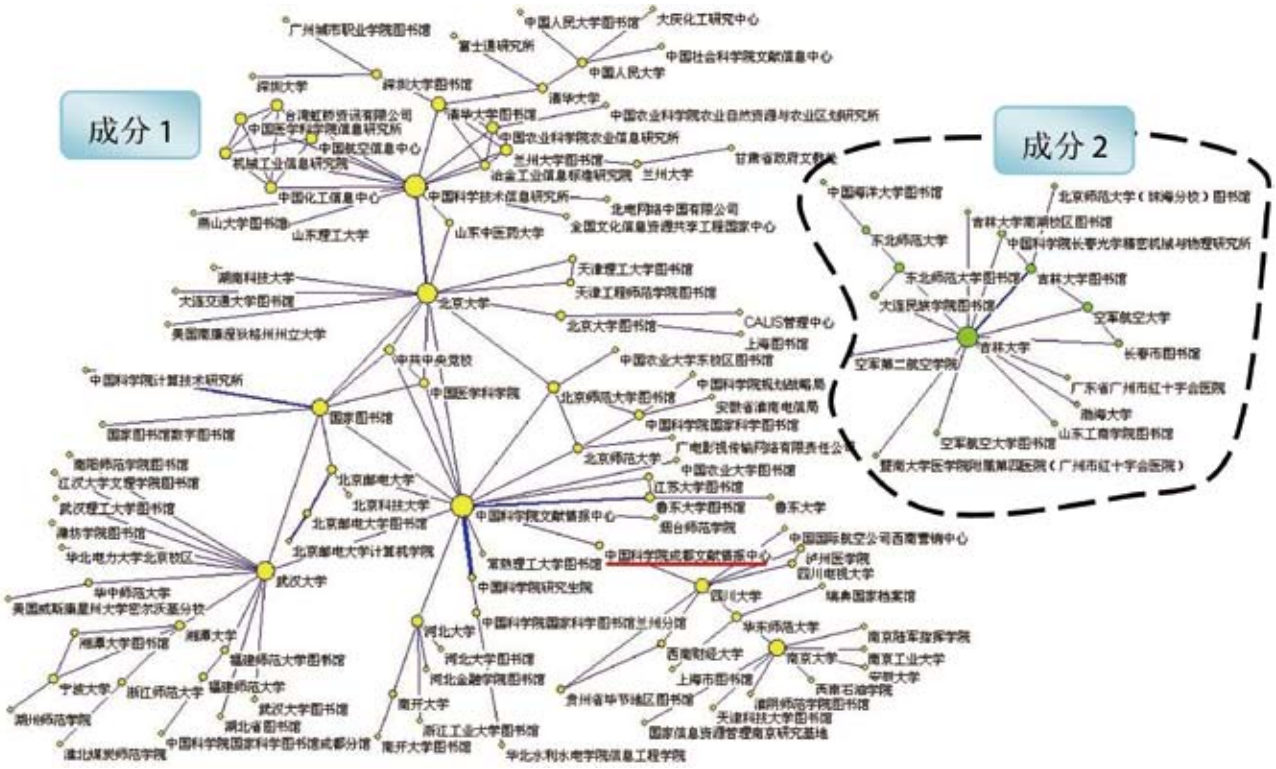

图1 机构合作网络中最大的两个成分

中的第二大成分，即成分2，这一成分中的节点与最大成分没有连接，说明近5年里，前者中的研究机构与后者中的研究机构在数字图书馆领域没有进行合作发文。该网络以是吉林大学为中心的机构合作网络，合作单位多为东北地区的研究机构，表现出较强的地域性倾向。

在表1中，中国科学院成都文献情报中心的介数排名第五，从图1中可以看到，虽然中国科学院成都文献情报中心只与四川大学和中国科学院文献情报中心两所研究机构合作发表文章，但它将四川大学、华东师范大学、南京大学等所在的分支网络与中国科学院文献情报中心所在的主干网络连在了一起，对整个网络的连通性有较大贡献。

3 我国数字图书馆领域作者合著网络分析

随着科学研究深度和广度的发展，各学科相互交叉渗透，产生许多新生领域，这些领域的研究需要多种学科的人才相互协作，联合攻关，因此，作者合著的论文也越来越多。数字图书馆领域涉及到的领域多、范围广，通过作者合作，充分发挥群体智慧，可以在知识结构等方面相互取长补短，提高研究成果的水平。

3.1 作者合著网络整体分析

对近5年国内数字图书馆作者合著的数据进行处理，得到相应的作者合著网络。其中，网络节点代表文章的作者，网络的边代表作者之间有共同署名发表的文章，边的权值代表作者之间共同发表文章的数量。作者合著网络的节点总数为2202个，边数为3436个，密度为0.000709，可见，研究者在数字图书馆领域中合作很少。

对作者合著网络中各个节点的度数和介数进行统计分析，得到度数和介数最高的前10位作者，如表3所示。根据度数的排名，与其他作者合著最多的是张晓林，接着是董慧、孙坦、张智雄和毛军等。根据介数的排名，在机构合作网络中处于中心位置的是张晓林，之后是毛军、张智雄、张晓青和孙坦等。其中，董慧度数排名较高，但未处于主干网络，因而介数排名不在前10名。

对作者合著网络进行成分划分，并根据每个成分所包含的节点数进行分类统计，如表4所示，共得到1105个成分，最大的成分只有124个节点，并且有1094个成分所包含的几点数不到10，整个作者合著网络非常松散，由此可见，数字图书馆领域中研究者之间的合作少，且多停留在小范围的合作，整个领域的合作研究有待进一步加强。

表3 度数和介数最高的前二十位作者

排名	度数	作者	介数	作者
1	19	张晓林	0.002135	张晓林
2	16	董慧	0.001216	毛军
3	15	孙坦	0.000767	张智雄
4	15	张智雄	0.000509	张晓青
5	15	毛军	0.000493	孙坦
6	14	胡铁军	0.000471	李书宁
7	12	张继东	0.000467	胡铁军
8	11	马建霞	0.000424	周晓光
9	11	姜赢	0.000347	吴振新
10	10	马自卫	0.000335	姜恩波

3.2 作者合著网络主要成分分析

由于作者合著网络非常松散，我们仅选取包含节点数最多的五个成分做进一步的分析，它们的节点数量分别为124、33、31、25、20。如图2所示，首先

表4 成分分类统计

节点数	1-9	10-19	20	25	31	33	124	总计数
成分频次	1094	6	1	1	1	1	1	1105

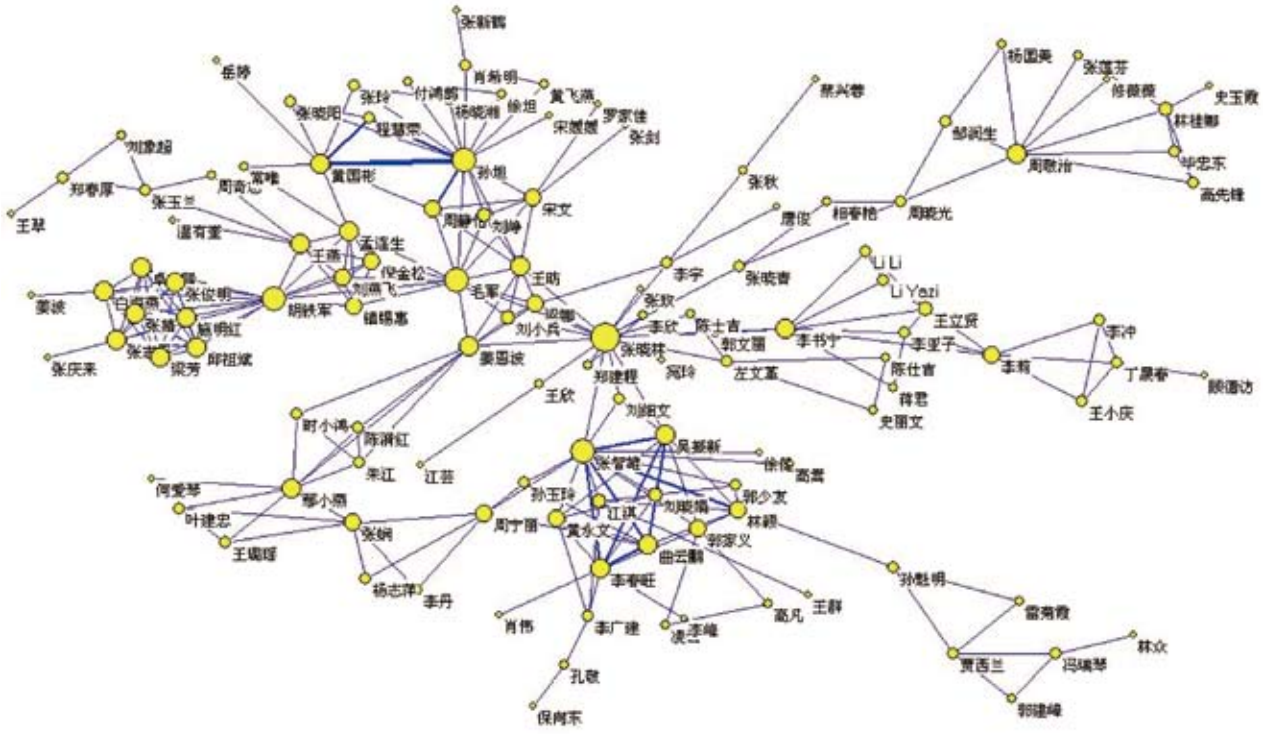

图2 作者合著网络中最大的成分

对作者合著网络中的最大成分进行单独分析，该网络的直径为11，平均路径长度为4.53。从图中可以明显看出，与其他作者合著较多的是张晓林、毛军、张智雄、孙坦和胡铁军，其中，处于较为中心位置的是张晓林，其次是毛军和张智雄，而孙坦和胡铁军则主要是在局部网络比较活跃，这与表3的介数排名一致。

在表3中，张晓青和李书宁的度数排名不在前20位，但他们的介数排名却分别为第4和第6。从图2中可以看到，虽然与张晓青和李书宁两个节点直接相连的边并不多，但这两个节点分别将两个较大的分支网络与主干网络连在了一起，对整个网络的连通性有较大贡献。

如图3所示，对作者合著网络中的其他四个较大的成分进行分析，成分2有33个节点，是第二大的网络社群，处于中心位置的作者是毕强、牟冬梅和韩毅，成分3有31个节点，处于中心位置的是董慧和马自卫，成分4有25个节点，处于中心位置的是孙卫和申晓娟，成分5有20个节点，处于中心位置的是董丽。这些成分中的核心作者虽然有较高的度数，但与主干网络没有连接，对整个网络的贡献相对较小，因此，对应的介数排名较低。

4 我国数字图书馆领域共词网络分析

共词分析可以概述学科研究热点，横向和纵向分析领域学科的发展过程、特点以及领域或学科之间的关系，反映某个专业的科学研究水平及其发展历史的动态和静态结构。这与高频关键词列表反映出来的领域热点有所不同，共词网络展现了数字图书馆研究中各个领域的划分，以及这些领域之间的相互联系。

4.1 共词网络整体分析

对近5年国内数字图书馆关键词共现的数据进行处理，得到相应的关键词共现网络，即共词网络。其中，网络节点代表文章的关键词，网络的边代表两个关键词出现在同一篇文章中，边的权值代表关键词共现的次数。共词网络的节点总数为3323个，边数为11270个，密度为0.00204，可见，共词网络比较松散，但相对机构合作与作者合著网络要更加紧密。

对关键词共现网络中各个节点的度数和介数进行统计分析，得到度数和介数最高的前20位关键词，如表5所示。除去“数字图书馆”、“图书馆”等意义

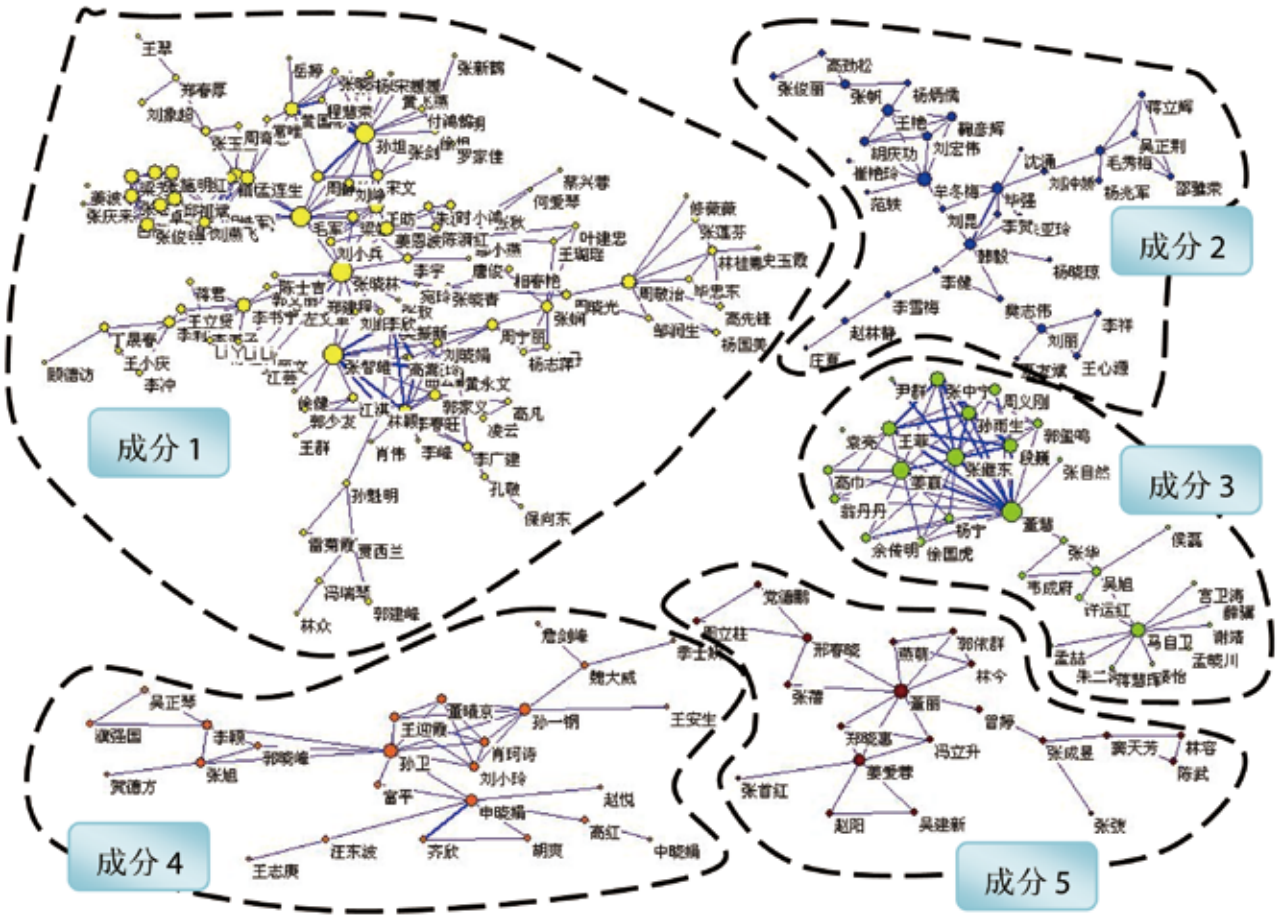

图3 作者合著网络中最大的五个成分

宽泛的词，根据度数的排名，与其他关键词共现频次较高的有元数据、本体、信息服务、数据资源、个性化服务、知识产权、著作权、信息资源和信息组织等词，这说明它们是数字图书馆中研究较多的方面；根据介数的排名，在共词网络中处于中心位置的词有本体、元数据、数字资源、网格、资源整合、个人数字图书馆和个性化服务等，这些词在整个数字图书馆领域中处于较为核心的位置。

对共词网络进行成分划分，并根据每个成分所包含的节点数进行分类统计，如表6所示，共得到82个成分，最大的成分有3097个节点，占总节点数的92.3%，整个共词网络的连通性很好，其他的成分最多只有6个节点，并且累计之和为226个节点，因此，可以将整个共词网络的分析简化为最大成分的分析。

4.2 共词网络主要成分分析

共词网络的最大成分包含3097个节点，该成分的直径为8，平均路径长度为2.82，有很好的连通性，节点之间的连接较为紧密。由于节点数量多，对其进行可视化后的显示效果不佳，我们对其进行简化，提取网络的主干部分，具体操作为：去掉了边值小于3的所有边，然后去掉节点“数字图书馆”和所有孤立点。简化后，选取包含节点数最多的三个子成分做进一步的分析，它们的节点数量分别为79、18、6。

如图4所示，黄色节点及其连线所构成网络是简化之后得到的最大成分，即成分1，有79个节点。从图中可以明显看出，与其他关键词共现较多的是图书馆、信息服务、信息组织、个性化服务、读者服务、元数据和本体等，其中，图书馆、信息服务、信息组织和个性化服务处于较为中心的位置，可见，在数字图书馆研究中非常重视为用户提供个性化的服务。从整个网络所形成的社群来看，研究主要集中在元数据、本体、个性化服务、数字版权、资源建设、推荐系统、

表5 度数和介数最高的前二十位关键词

排名	度数	关键词	介数	关键词
1	1871	数字图书馆	0.76720	数字图书馆
2	199	图书馆	0.04632	图书馆
3	107	数字图书馆建设	0.01714	本体
4	101	元数据	0.01377	数字图书馆建设
5	98	本体	0.01355	元数据
6	95	信息服务	0.01313	高校图书馆
7	95	数字资源	0.01038	数字资源
8	84	个性化服务	0.01035	数字图书馆系统
9	84	国家图书馆	0.01001	图书馆联盟
10	80	高校图书馆	0.00973	网格
11	74	数字化	0.00937	资源整合
12	68	应用	0.00915	信息服务
13	67	知识产权	0.00910	个人数字图书馆
14	66	数字化图书馆	0.00905	个性化服务
15	60	图书馆联盟	0.00893	层次分析
16	59	著作权	0.00881	数字化图书馆
17	59	数字图书馆系统	0.00824	复合图书馆
18	57	体系结构	0.00805	数字化
19	52	信息资源	0.00770	图书馆服务
20	52	信息组织	0.00727	向量空间模型

表6 成分分类统计

节点数	1	2	3	4	5	6	3097	总计数
成分频次	27	8	14	21	9	2	1	82

管理创新等方面，其中，个性化服务与个人数字图书馆、本体、信息组织、信息服务、推荐系统等联系比较紧密，而其他研究之间的联系很少，在社群网络中比较独立，如本体领域的研究只与个性化服务有关联。此外，每个研究主题也有不同的侧重，如信息组织领域中的元数据是研究比较多的方面。

成分2有18个节点，主要是图书馆管理和技术领域，关键词之间共现频率高，联系紧密。成分3有6个节点，主要涉及信息存储领域。

5 总结

通过对数字图书馆领域机构合作、作者合著与关键词共现三个网络的分析，整体上描述了2005-2009年数字图书馆领域研究的状况，得到了处于核心位置的研究机构和作者，并且分析结果显示，数字图书馆领域的机构合作与作者合著较少，有待加强。对数字图书馆关键词共现网络的分析显示，研究热点主要集中在元数据、本体、个性化服务、数字版权和资源建设等方面。

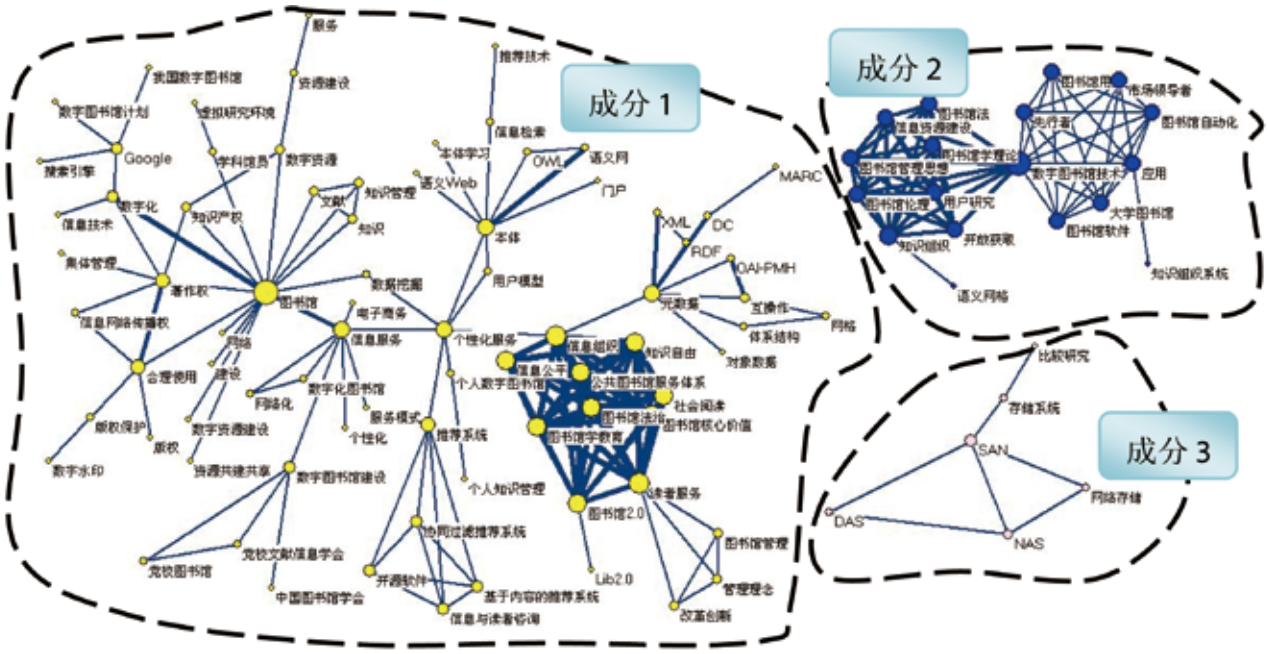

图4 共词网络的三个最大子成分

参考文献

[1]李春.关注网络社群[J].思想·理论·教育,2004(12):9-13.
[2]林聚任.复杂网络理论及其研究[M].北京:清华大学出版社,2005.
[3]裴雷,马费成.社会网络分析在情报学中的应用和发展[J].图书馆论坛,2006,26(6):40-45.
[4]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,2006,25(5):634-640.
[5]苏娜.基于共词分析的数字图书馆领域研究主题及进展分析[J].情报学报,2009,28(6):15-19.
[6]王福生,杨洪勇.《情报学报》作者科研合作网络及其分析[J].情报学报,2007,26(5):659-663.
[7]李亮,朱庆华.社会网络分析方法在合著分析中的实证研究[J].情报科学,2008,26(4):549-555.
[8]邱均平,马瑞敏,李晔君.关于共被引分析方法的再认识和再思考[J].情报学报,2008,27(1):69-74.
[9]林聚任. 社会网络分析：理论、方法与应用[M].北京:北京师范大学出版社2009.

作者简介

张鹏（1984-），北京大学信息管理系硕士研究生，研究方向：信息检索与分析。通讯地址：北京大学信息管理系 100871。E-amil：pengzhang08@gmail.com
王建冬（1982-），北京大学信息管理系博士研究生，研究方向：信息服务与情报分析。通讯地址：同上
王继民（1966-），北京大学信息管理系副教授，博士，研究方向：搜索引擎与Web挖掘。通讯地址：同上

A Statistical Analysis of the Papers on Digital Library Research in China （2005-2009） --- Community Analysis

Zhang Peng, Wang Jimin, Wang Jiandong / Department of Information Management, Peking University, Beijing, 100871

Abstract: Based on digital library research literature from Wanfang Data between 2005 and 2009, we analyze institutional cooperation, co-author and co-word networks, and get the three community distribution maps. Then, we make the evaluation of core institutions and authors, and conduct a sub-field of research and outline the research priorities. Keywords analysis shows that digital library research focus on metadata, ontology, personalized services, digital rights , resource development and organization and so on.

Keywords: Digital library, Social network analysis, Institutional cooperation, Co-author network, Co-word network

（收稿日期：2010-01-14）