基于CIT系统的数字图书馆学科的热点研究

□ 朱希田 / 吉林化工学院图书馆 吉林 132000

摘要：文章的研究对象为高校优秀硕、博学位论文，研究方法为引文共引聚类分析与关键词词频分析相结合的方法。根据引文共引理论设计了CIT系统。通过操作CIT系统得出相关矩阵，将CIT系统计算的数据导入SPSS中聚类，得出了客观的数据结果。最后，科学地分析了国内外数字图书馆学科的研究热点，并进行了比较。

关键词：数字图书馆，学位论文，共引聚类，词频分析，CIT系统
DOI：10.3722/j.issn.1673—2286.2009.04.016

当前，数字图书馆的建设成为了评价国家基础建设的一项重要指标。数字图书馆的建设离不开数字图书馆学科的理论指导，推进数字图书馆学科的理论研究并不是盲目的专研，而是要依据近年来数字图书馆学科发展情况有方向有重点地推进。对数字图书馆学科热点的分析有助于整个数字图书馆的合理建设。

1 本文的研究对象和研究方法的确立

本文的研究对象是高校优秀的硕博学位论文。作者认为研究数字图书馆学科状态最重要的体现之一是高校对数字图书馆的研究，尤其是对高校优秀硕士、博士学位论文的研究。因为从学位论文的角度看，国内外教育系统中对硕士、博士的研究培养是目前国际上教育体系中最高层次的教育，代表着一个学科的进展程度。[1]

具体的研究数据收集如下：国内数据库选取的数据源是CNKI优秀硕、博学位论文数据库，它是目前国内相关资源最完备、高质量、连续动态更新的学位论文全文数据库[2]。检索条件是：关键词或题目中含有“数字图书馆”字样的学位论文，检索时间为2007年6月12日，共检索出272条符合检索条件的论文。保存这272篇.pdf格式的论文，利用Office 2003 插件ScanSoft PDF Converter for Microsoft Word v4.0将 272篇论文的引文部分的.pdf格式转换成.doc格式。最后，将转换的.doc格式的文档导入到SQL Server2000数据库中即可。国外数据库选取的数据源是PQDD，因为ProQuest公司是世界上最早及最大的博硕士论文收藏和供应商。该公司收集有170万篇国外高校硕博士学位论文[3]。检索条件是：论文关键词或者题目中含有“digital library”字样的学位论文，检索的时间为2007年6月12日，共检索出15条符合检索条件的论文，收集数据的处理同中文数据库所述。

本文的研究方法是引文共引聚类分析与关键词词频分析相结合的研究方法。这种方法是近年研究引证关系和文献微观结构的一种最新方法。

共引：如果两篇论文A和B同时被后来的一篇或者多篇论文所引用，则认为文献A和B具有共引关系，同时引用A和B的论文数量为共引强度。强度越大，文献A和B越相关[4]。聚类：本文的聚类方法是快速聚类，将数据看成是k维空间上的点，以个体间距离作为测度聚类的指标[5]。关键词词频分析主要是对关键词的出现频率进行统计。如果该词频出现较高、较稳定，则说明对该热点的研究近几年较稳定，有可能在未来几年仍然是研究的热点；如果出现前些年该词频较高，近两年词频下降的情况，则说明该关键词的热点在降温，很可能在未来几年中淡出[6]。

2 CIT系统的实现

通过对系统的分析，根据Jackson设计方法（Jackson设计方法是计算机软件工程中用来设计总体设计图的主要方法）设计出了总体设计图（图1）。

CIT系统的前台设计工具是PB9.0，后台数据库工具是SQL

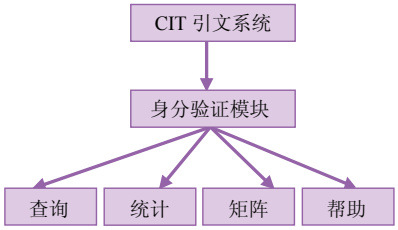

图1 CIT系统的总体设计图

Server2000。具体的连接界面和功能主界面如图2、图3所示。

系统的主要的功能是通过菜单或子菜单控件来实现的，如文献关键词的查找和统计、高频引文的统计、计算共引矩阵、形成相关矩阵等。

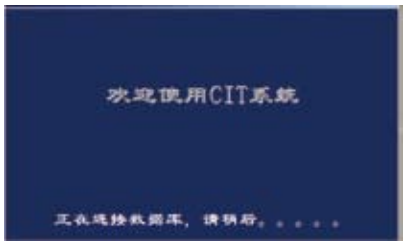

图2 连接界面

图3 功能主界面

3 基于CIT系统对国内数字图书馆学科热点的分析

经操作CIT系统得出相关矩阵，见表1（数列共37列显示不下，只列出了前10列，并将此对称矩阵补全了下三角）：

将此结果复制到SPSS中通过K-Means的聚类方法进行聚类，如图4。

表1 相关矩阵表

0	1	2	3	4	5	6	7	8	9	10
1	1	0.33	0.22	0.30	0.05	0	0.05	0.05	0	0.12
2	0.33	1	0.38	0.83	0	0	0	0.25	0.11	0.06
3	0.22	0.38	1	0.25	0	0	0	0.29	0.06	0.06
4	0.30	0.83	0.25	1	0	0	0	0.29	0.13	0
5	0.05	0	0	0	1	0	0.13	0	0	0
6	0	0	0	0	0	1	0	0	0.06	0
7	0.05	0	0	0	0.13	0	1	0	0	0
8	0.05	0.25	0.29	0	0	0	0	1	0.14	0.07
9	0	0.11	0.06	0.13	0	0.06	0	0.14	1	0
10	0.12	0.06	0.06	0	0	0	0	0.07	0	1

选择菜单【Analyze】-【Classify】-【K-Means Cluster】，进入到下图的参数设置界面如图5。

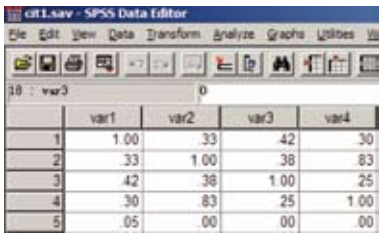

图4 数据复制到SPSS中

图5 聚类的参数设置界面

在“number of clusters:”选项下输入5，表示想将37篇高频引文聚成5类。在“variables”中定义37个变量，即将1到37列定义成37个变量名，方便聚类时说明每一个类下的类成员名称。然后点击“OK”。所有的聚类工作都由SPSS软件系统自动完成，无需人工操作。

聚合后的各类论文，通过模糊检索发现这几类关键词出现的频率较高：资源、服务检索技术、参考咨询、著作权类等。最后借助CIT系统的关键词查询功能统计出每一类各关键词出现的频次，其中关键词总计582个。以下是通过统计词频高低得出的各类热点问题。

3.1 含有“服务”的各高频关键词

服务类的关键词共计108个。其中“个性化服务”、“知识服务”、“服务质量”的数量所占比例较大。服务类关键词最早出现在2001年，研究贯穿于各个年段并呈逐年上升的趋势，主要集中在个性化信息服务、知识服务、服务质量方面的研究。

3.2 含有“资源”的各高频关键词

表2 聚类结果

Case Number	Cluster Number of Case	Cluster	Distance
1	2	2	.790
2	2	2	.650
3	2	2	.754
4	2	2	.690
5	5	5	.779
6	4	4	.977
7	1	1	.958
8	1	1	.943
9	1	1	.939
10	3	3	.646
11	3	3	.684
12	2	2	1.047
13	3	3	.863
14	3	3	.801
15	2	2	.884
16	1	1	.946
17	1	1	.825
18	4	4	.890
19	5	5	.423
20	4	4	.878
21	3	3	1.123
22	5	5	.453
23	4	4	.938
24	4	4	.859
25	5	5	.505
26	5	5	.451
27	1	1	.982
28	3	3	.802
29	3	3	.811
30	1	1	.885
31	4	4	.976
32	4	4	.943
33	3	3	.771
34	1	1	.844
35	5	5	.746
36	4	4	.904
37	5	5	.432

表3 服务类关键词的各年分布情况

年份 / 数量	2000	2001	2002	2003	2004	2005	2006	总计	比例
个性化服务	0	3	3	6	3	9	8	32	30%
知识服务	0	0	0	4	7	9	10	30.	28%
服务质量	0	0	0	0	0	4	8	12	11%

资源类关键词共计96个。最早出现于2001年，从2002年开始增长较快，从关键词词频统计上看数量上位于前三位的是“资源整合”、“资源组织”、“资源管理”。

表4 资源类关键词的各年分布情况

年份 / 数量	2000	2001	2002	2003	2004	2005	2006	总计	比例
资源整合	0	0	0	0	4	8	14	26	27%
资源组织	0	2	1	5	7	4	6	25	26%
资源管理	0	0	1	3	4	8	7	23	24%

3.3 含有“参考咨询”的各高频关键词

参考咨询类关键词共计89个。最早出现于2002年，增长较快，词频数量位于前三位的是“数字参考咨询”、“合作式参考咨询”、“参考咨询服务质量及标准”。

表5 参考咨询类关键词的各年分布情况

年份 / 数量	2000	2001	2002	2003	2004	2005	2006	总计	比例
数字参考咨询	0	0	1	4	7	12	13	37	42%
合作式参考咨询	0	0	0	0	3	6	11	20	22%
参考咨询质量、标准	0	0	0	1	2	4	6	13	15%

3.4 含有“检索”的各高频关键词

检索类关键词共计68个。最早出现在2001年。从统计数据上看研究年代具有连续性，但是研究点比较分散。“集成检索”17次，“全文检索”7次，“信息检索”出现5次，“多媒体检索”4次等（“图像检索”，“音频检索”共出现4次，将图像检索，音频检索统称为多媒体检索）。值得关注的是，在检索这一类论文关键词中经常伴有“元数据“、“XML”、“OAI”等关键词字样，经统计“元数据”共出现41次,“XML”共出现25次，“OAI-PMH”共出现21次。因此将元数据、

XML、OAI-PMH归并为检索类。从统计数据看检索领域正朝着新技术领域发展。关键词“元数据”、“XML”、“OAI-PMH”与检索类关键词数计算后合计为129次。

表6 检索类关键词的各年分布情况

数量 \ 年份	2000	2001	2002	2003	2004	2005	2006	总计	比例
元数据	0	7	4	5	11	6	8	41	32%
XML	0	4	2	3	8	5	3	25	19%
OAI_PMH	0	0	1	2	9	4	5	21	16%
集成检索	0	0	0	2	4	6	5	17	13%

3.5 含有“著作权、知识产权、数字版权”的各高频关键词

著作权类关键词共计49个。最早出现在2002年，2004-2006出现次数较多，增长也较快，且以著作权出现频率最高，共出现27次。

表7 著作权类关键词的各年分布情况

数量 \ 年份	2000	2001	2002	2003	2004	2005	2006	总计	比例
著作权	0	0	0	3	4	4	16	27	55%
知识产权	0	0	0	0	2	2	3	7	14%
数字版权	0	0	2	0	2	0	1	5	10%

4 基于CIT系统对国外数字图书馆学科热点的分析

国外的数据统计及计算与国内研究的思路一致，因此得出近年来国外数字图书馆学科主要研究方向有服务、资源、检索三大方面。国外的研究时间比国内的早，国外最早时间是1983年。经检索，符合外文查询条件的源文献只有15篇，对应的参考文献共324篇，高频引文对应的关键词总计72个，数据量不是很大，因此关键词统计的总量及各年的分布都比较少，从数量上说明不了问题。但是从相对值的角度，即所占比例上判断其研究重点即可。服务类关键词约占36%，资源类约占30%，检索类约占26%。

4.1 含有“server”的各高频关键词

服务类关键词共计26个。国外主要集中在数字图书馆的检索模式服务、检索结果两方面的研究。

表8 外文文献中server类关键词的各年分布情况

数量 \ 年份	Model of Seeking server	Result of research server
2000	1	1
2001	1	1
2002	1	0
2003	3	0
2004	0	1
2005	3	1
2006	1	2
总计	10	6
比例	38%	23%

4.2 含有“resource”的各高频关键词

资源类关键词共计21个。通过数据统计分析看，国外近年对资源的研究主要集中在资源评价和资源标识上。

4.3 含有“retrieve”的各高频关键词

检索类关键词共计19个。通过数据统计分析看，国外近年对检索的研究主要集中在搜索引擎和数据挖掘两个方面的研究。

5 对国内外分析的结果比较

从统计数据的时间上看，对数字图书馆的研究国外的时间比国内的时间早，国外最早的时间是1983年，国内最早的时间是2000年。

从统计的关键词数量变化上看，国内数字图书馆论文中关键词“服务”和“资源整合”占据数量的绝对优势，对著作权等领域的研

表9 外文文献中 resource类关键词的各年分布情况

年份/数量	2000	2001	2002	2003	2004	2005	2006	总计	比例
identify resource	1	1	2	0	0	2	1	7	33%
Resource marking	0	1	1	1	0	0	2	5	24%

表10 外文文献中retrieve类关键词的各年分布情况

年份/数量	2000	2001	2002	2003	2004	2005	2006	总计	比例
Search engine	1	1	0	0	1	1	2	6	32%
Data mining	0	1	1	0	1	1	1	5	26%

究数量也在逐年递增。说明我国数字图书馆正向着资源服务的深度（如个性化服务）和资源服务的广度（如资源整合）方向上发展。此外，对数字图书馆发展的规范化研究（如著作权）也逐渐成为重中之重。国外的数字图书馆论文中关键词“服务”、“资源”占数字图书馆学科总量上的研究比例最大，可是与国内比较看国外更侧重从小的方面进行集中研究（如对检索结果的服务）。

从发展趋势上看，国内的研究重点集中在服务、资源、参考咨询、检索、知识产权、著作权等研究领域上。从各领域研究的重心看目前国内数字图书馆主要侧重于数字图书馆的个性化服务理念、参考咨询服务模式、数字资源的整合、著作权等方面的研究上。国外发展的主要领域是服务模式（如分类式交谈系统）、资源标识、评价以及搜索引擎、数据挖掘在图书馆中的具体应用等。另外，国外在服务模式上的研究很注重对用户行为理念的跟踪调查。国外在图书馆资源方向上的研究已经跨越了资源整合阶段，已步入到资源评价的层面上。在检索技术上国内外对数据挖掘等检索技术进行深入地研究，并更加注重检索的效率问题。值得注意的是近年国外在检索技术上比较侧重根据用户行为习惯对检索结果二次处理方面的研究。对比来看国内的研究范围比较广，但多为理论上的概述；国外的研究范围较小，但重点突出，研究相对比较深入。

综上所述，国内外研究的起始时间不同，并且从发展的重点看，国内外在大的方向上保持一致，都是集中在服务和资源上的研究。但是，在大方向下的研究却各有侧重，希望国内能够借鉴国外好的发展模式和理念来扩充国内的研究，使数字图书馆在国内的社会发展建设中起到更加重要的作用，更好地为社会、为人民服务，体现图书馆的重要价值。

参考文献

[1] 叶琦. 博士学位论文引文分析与研究[J] . 中华医学图书情报杂志,2005,14(5):62-64.
[2] 清华同方中国优秀博硕士学位论文全文数据库[OL].[2007-06-12]. http://dlib.cnki.net.
[3] 国外学位论文数据库ProQuest Digital Dissertation[OL]. [2007-06-12].http://proquest.calis.edu.cn/umi/index.jsp.
[4] 陈定权.同引分析与可视化技术[J].情报科学,2005,23(4):531-537.
[5] 薛微.基于SPSS的数据分析[M].北京: 中国人民大学出版社,2006:328-334.
[6] 吴稔年.2000-2004年中国图书馆事业热点分析[J].新世纪图书馆,2006(1):25-26,30.

作者简介

朱希田（1957-），男，毕业于北京科技大学，现为吉林化工学院图书馆馆长，吉林省图书馆学会理事，吉林省高校图书工作委员会委员，副研究员。通讯地址：吉林化工学院图书馆，132000。E-mail: zxttsg@sina.com

The Hotspot Research of Digital Library Disciplines based on CIT System

Zhu Xitian / Jilin Chemical Technology Institute Library, Jilin, 132000

Abstract: This passage uses the excellence dissertations and theses databases as research object, the method is co-citation cluster analysis and keywords frequency analysis. Using co-citation theory, CIT system is designed. Using the CIT system, it gets the correlation matrix and puts the data into SPSS tool to cluster, then it gets the impersonal data. According to the data, it analyzes the hotspots of digital library research both at home and abroad.

Keywords: Digital library, Dissertation, Co-citation cluster, Keywords frequency, CIT system

（收高日期：2008-08-16；责任编辑：贾延霞）