第 28 卷 第 6 期
2009 年 6 月
情 报 杂 志
JOURNAL OF INTELLIGENCE
Vol. 28 No. 6
June 2009

基于共词分析的数字图书馆领域研究主题及进展分析

Analysis of Subjects and Development in Digital Libraries Research Based on Co - words Analysis Method

苏 娜

(中国科学院国家科学图书馆 北京 100190;中国科学院研究生院 北京 100049)

摘 要 以 ISI 的 Web of Science 数据库为数据来源,采用共词方法分析数字图书馆领域的研究主题和研究进展及趋势。分析结果表明,数字图书馆领域的研究主题主要为数字图书馆信息组织和描述标准研究、数字图书馆技术研究两个部分。目前,在信息组织和描述标准方面,元数据和本体依然是研究的热点;而在技术方面,除了信息检索这个贯穿数字图书馆领域的研究内容,用户界面和用户研究成为近两年来的热点。此外,对学习对象和数字仓储的研究、数字图书馆教育和评价研究也成为近两年来研究者关注的热点。

关键词 数字图书馆 共词分析 研究主题

中图分类号 G250.76 **文献标识码** A **文章编号** 1002 - 1965(2009)06 - 0015 - 05

数字图书馆的研究始于 20 世纪 80 年代末 90 年代初。1993 年美国国家科学基金会(NSF)、美国国防部高级研究计划署(DARPA)、国家航空航天署(NASA)联合发起数字图书馆创始工程(Digital Library Initiative, DLI),开启了数字图书馆规模建设的大幕,此后十几年中,世界各国开始进行大规模的数字图书馆建设,数字图书馆的研究随之成为图书情报和计算机等信息科学领域的研究热点,研究成果不断丰富,研究内容不断深化。

数字图书馆是一项多学科集成的综合系统工程,研究内容广泛而多角度,采用定性的方法对数字图书馆研究进行全景分析和综述需要研究者具有全面的数字图书馆领域知识和专业经验,对研究者要求较高,并且带有很强的主观色彩,因此研究者往往采用定量化的研究方法来分析该领域的研究现状。如晏尔伽、朱庆华以美国科学情报研究所开发的 SCI - E 数据库为数据源基础,从文献量、著者、机构、核心期刊、引文等角度进行了统计和分析,以定量数据从侧面反映了近十年来数字图书馆的发展情况[1]。周静怡、孙坦以 web of knowledge 数据库中收录的 1993～2004 年间有关数字图书馆领域的文章为数据来源,从论文的发表时间分布、期刊分布、被引频次分布、作者分布四个方面进行了统计与分析,初步确定了数字图书馆领域的核心期刊、经典文献和核心作者[2]。赵秀君对 1994～2003 这 10 年间刊载在我国 15 种图书情报学核心期刊上的有关数字图书馆研究的论文进行了定量分析,分析的内容具体包括各年度发文情况、论文产出期刊源及其分布、论文及作者的地区分布和系统分布、作者人均发文量、核心作者的地域分布及其论文的主题分布等[3]。此外,钟云志、周东晓、杜香莉等人通过建立我国数字图书馆论文与作者数量关系的洛特卡定律来揭示数字图书馆论文——作者的数量分布[4]。这些研究的一个共同特点是对文献数量、著者、机构、核心期刊、经典文献等内容进行计量分析,多角度定量描述数字图书馆领域的研究现状。在此基础上,晏尔伽、朱庆华以 LISA 为数据来源,采用词干网方法对有关数字图书馆的研究文献的叙词进行网络化处理,以定量数据从侧面反映数字图书馆研究的发展情况[5]。

上述提到的数字图书馆定量研究主要集中在对 2005 年以前的数据的分析,而且多是利用频次进行统计描述,对最新的研究发展情况尚无太多的研究。本文将以上述研究为基础,在对 1990 年以来数字图书馆研究主题进行整体分析的基础上,通过对最新数据的分析,进一步描述数字图书馆研究中的各研究主题。

收稿日期:2008 - 11 - 20 修回日期:2009 - 01 - 13

作者简介:苏 娜,女,1983 年生,博士研究生,研究方向为学科情报与战略情报研究。

除了利用频次统计的方法外,本文还将通过共词和可视化的方法进一步研究主题的变化情况。

1 本文所采用的方法和数据来源

共词分析法属于内容分析法的一种,它的原理主要是对一组词两两统计它们在同一篇文献中出现的次数,以此为基础对这些词进行聚类分析,从而反映出这些词之间的亲疏关系,进而分析这些词所代表的学科和主题的结构变化[6]。

早在20世纪70年代中后期法国计量学家就对共词分析方法进行了详细的描述。共词分析方法第一次被应用是在20世纪80年代的法国,通过一个称为"LEXIMAPPE"的系统实现。1986年Callion在他的著作《科学技术动力学的绘图》中进一步介绍了共词分析方法[7]。经过将近30年的发展,共词分析方法已经被广泛应用到许多领域,主要集中在人工智能、科学计量学、人文学科计算研究、信息科学和信息系统的研究、信息检索等领域。基于此原因,本文采用共词的方法来分析数字图书馆领域的研究主题和发展趋势。

美国科学信息研究所ISI的Web of Science数据库是世界上著名的网络数据库,包括SCI、SSCI和A&HCI三个索引数据库。Web of Science数据库的选刊标准和评估程序非常严格,从而能够保证其收录的文献能涵盖全世界最重要和最有影响力的研究成果。因此本文以web of science数据库为数据来源,以"digital library"为主题词进行检索。将文献类型限定为"article"和"proceedings paper",共得到2977篇文献,去重后精简为2968篇。在对数据进行整理和聚类分析的过程中,主要采用Excel对数据进行统计分析,利用Ucinet对共词矩阵进行网络指标分析和可视化。

2 结果分析

2.1 数字图书馆领域论文数量的变化情况及分析

发文数量的变化情况可以清晰地描述某一领域研究的不同发展阶段。图1描述了数字图书馆领域文章发表数量随年份变化的情况。数字图书馆领域的论文20世纪90年代初开始出现,在1995年之前经历了缓慢的发展,发文数量均低于50篇;1995年之后数字图书馆的研究成果迅速增长,到2004年达到顶点,尤其在2002年到2004年间有关数字图书馆的研究论文成直线增长趋势;2004年以后论文发表数量开始下降。图1大体描绘了数字图书馆领域研究的发展变化情况,1995年之前为研究初期阶段,数字图书馆的研究开始出现并缓慢发展;1995年之后随着世界各国数字图书馆建设的大规模开展,数字图书馆研究进入高速发展期;到2004年发文量达到顶峰,之后论文数量开始下降。这一方面可能是由于数字图书馆领域研究日趋成熟,其研究内容更加深化;另一方面是随着研究的不断成熟,研究者开始将视线转移到其他交叉研究领域。由于2008年的数据只是一部分,并没有包括全部的数据,所以2007年到2008年论文数量的骤减和样本的选取有关,并不代表真实的情况。

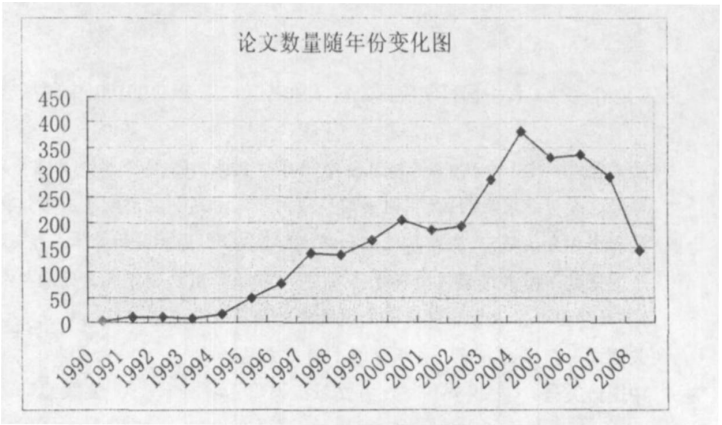

图1 论文数量随年份变化图

2.2 词频统计及分析

关键词是作者对文章的高度概括,统计分析关键词,能够由此发现数字图书馆领域的研究主题和热点。在本文所构建的数据集中,作者给定的关键词共4075个,经过清洗去重,得到3873个。表1为出现10次以上的高频关键词。

表1 部分关键词及出现次数列表

关键词	出现次数	关键词	出现次数
digital library	377	archive	12
metadata	43	digital	12
information retrieval	34	web services	12
internet	33	collaboration	11
libraries	29	digital images	11
ontology	21	digital video library	11
evaluation	20	e - learning	11
visualization	18	knowledge management	11
interoperability	16	segmentation	11
World Wide Web	16	user studies	11
education	14	annotation	10
Database	14	digital signal processing	10
educational digital libraries	14	FPGA	10
information visualization	14	indexing	10
XML	14	MPEG - 7	10
image processing	13	multimedia	10
user interfaces	13	personalization	10
architecture	12	usability	10

由于在数据集构建过程中采用的主题词为"digital library",出现的频次过多,因此在研究现状、热点和趋势描述时没有太大的意义,所以出现次数前10的关键词依次为:metadata、information retrieval、internet、li-

braries、ontology、evaluation、visualization、interoperability、World Wide Web、education(在后面的共词分析中依然将关键词 digital library 去除)。这些关键词代表了数字图书馆研究的重点主题领域。图 2 为这些关键词随年份出现的曲线图,通过它可以进一步描述这些研究主题在不同时期的分布,分析数字图书馆研究领域不同时间的研究重点。通过图 2,我们可以得出如下结论:

a. 信息检索的研究贯穿数字图书馆研究的整个过程,成为该领域的一个重要研究主题。

b. 在网络环境这个大背景下对数字图书馆的研究是 2003 年以前数字图书馆研究领域的重要内容。

c. 元数据的研究是数字图书馆领域一个最突出的研究热点。尽管 90 年代对元数据的研究已经开始出现,但从 2003 年开始元数据的研究才成为研究的热点,到 2005 年达到顶峰。对互操作的研究在 2003 年开始出现,随元数据一同发展,是元数据研究中重要的组成部分。

d. 对本体、数字图书馆教育、数字图书馆评价、可视化的研究成为近几年来数字图书馆领域的研究热点。

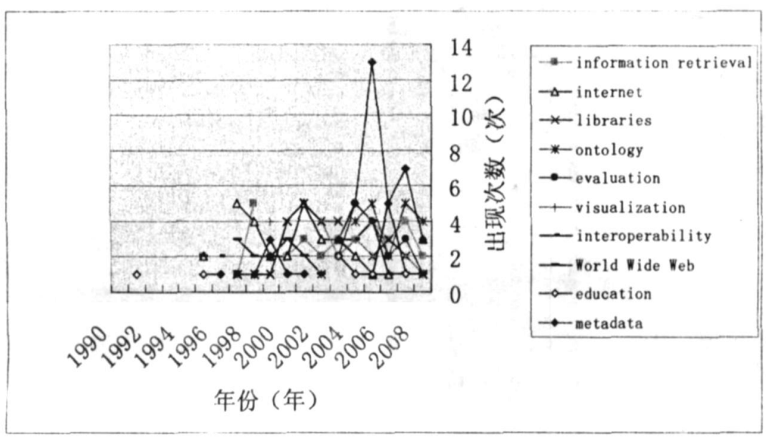

图 2　排在前 10 位的关键词随年份变化图

2.3　**数字图书馆领域研究主题分析**　为了有效地对研究领域的结构进行分析,本文选取出现次数为 5 次以上的 115 个关键词构建共现矩阵,然后利用 Ucinet 软件对共词矩阵进行可视化,如图 3 所示。图中节点的大小代表不同的度(Degree),是指与该节点相连的线的条数;不同的颜色代表节点间的中介中心性(Betweenness),是指一个节点处于其它节点最短路径上的个数;节点间线段的粗细则代表了节点间的关联强度。

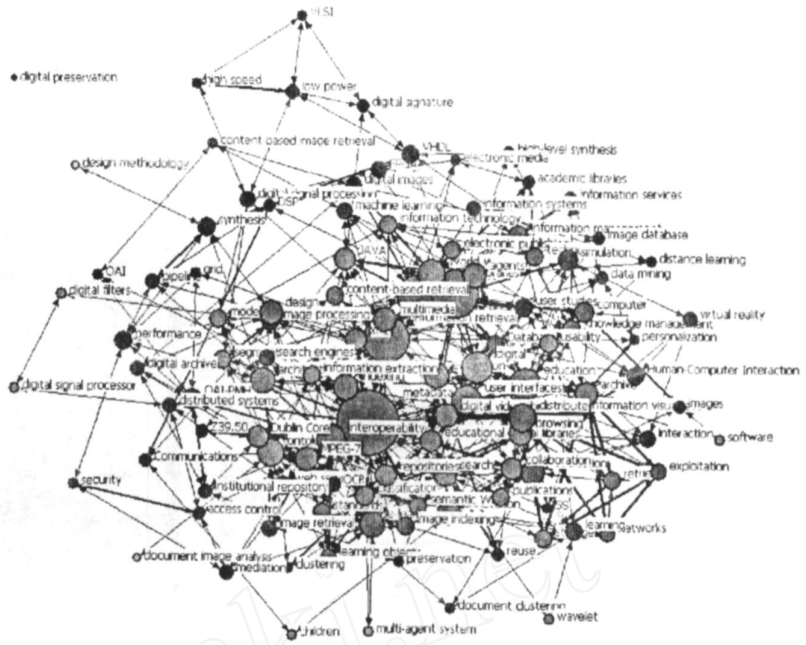

图 3　数字图书馆领域关键词共现网络

为了更加清晰地描述数字图书馆领域的主要研究主题,本文对共现次数低于 2 的关系进行了屏蔽,即将共现矩阵中低于 2 的值赋 0 值,得出数字图书馆领域的研究主题聚类图,如图 4 所示(不包括度数为 0 的节点)。

从图 4 中可以看出,1990 年到 2008 年间数字图书馆领域的研究虽然涉及到众多方面,但总体可以归纳为四个部分:区域 A 为数字图书馆信息组织标准研究,区域 B 为网络环境下数字图书馆技术研究,区域 C 为数字图书馆建设的流程研究,区域 D 则描述了数字图书馆研究的其他方面。这其中,区域 A 和区域 B 是数字图书馆研究的重点部分,包含诸多研究内容,而这些内容围绕着中心节点构成网状网络,节点之间又彼此相互关联,聚集成各自的研究簇。区域 C 单独成为一个研究点(可能是因为屏蔽造成的),而 A、B 和 D 三个区域则彼此关联。

a. 对数字图书馆信息组织标准的研究主要集中在元数据的研究,包括都柏林核心元数据标准研究、元数据的互操作研究、xml、多媒体内容描述接口、OAI - PHI 元数据收割标准、图像标引等。除此之外,对 Z39.50 和本体的研究也是数字图书馆信息组织标准

研究的重要方面。

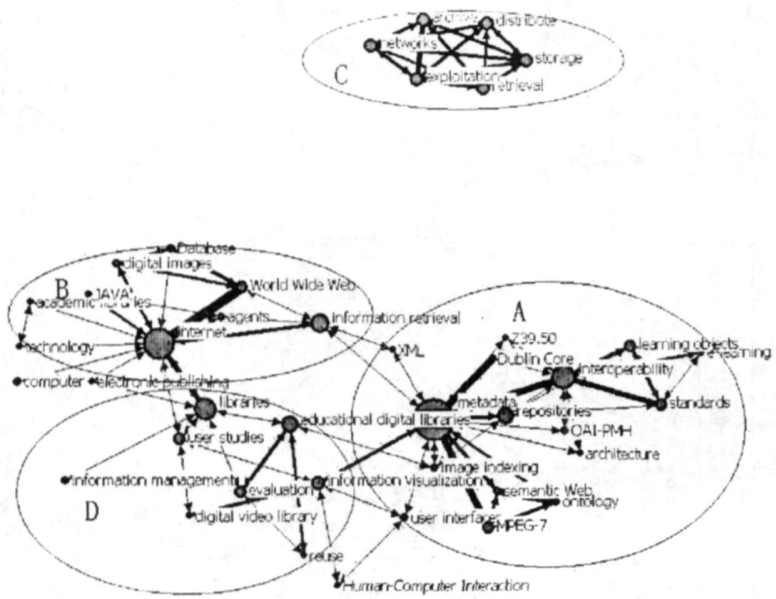

图4 经过处理后的关键词共现网络

b. 数字图书馆技术的研究主要围绕网络环境下数字图书馆的技术支撑展开,涉及到信息检索技术、数据库技术、编程语言、计算机、数字图像等。信息检索是数字图书馆技术研究中一个长盛不衰的子主题,对信息检索的研究主要集中在信息抽取和搜索引擎两方面。

c. 区域C描述了数字图书馆的流程,包括存档、分布、保存、检索、开发等几部分。数字图书馆流程的几方面相互关联度很高,密不可分。

d. 区域D是数字图书馆领域研究中相对较为分散的一个区域,研究内容涉及到具体类型数字图书馆的研究(主要集中在数字视频图书馆和教育数字图书馆)、数字图书馆的评价研究、信息可视化、用户界面和用户研究、人机交互等几个方面,还出现了信息管理、知识管理、信息服务等研究内容。这些研究相比较于其他三个区域,关联度并不高,研究内容比较分散。

通过上述分析可知,数字图书馆研究的主题主要为信息的组织和描述标准、数字图书馆技术两大部分。而对数字图书馆信息服务的研究还未成为研究的重点领域。

2.4 **数字图书馆领域研究进展分析** 上一节对1990年以来数字图书馆领域研究的整体进行了分析,指出数字图书馆的研究主要集中在信息组织和描述标准与数字图书馆技术两大主题,为了进一步深入分析研究数字图书馆的研究现状,进而发现该领域的研究趋势,笔者选取了2006～2008年的772篇文献构建数据集,利用共词分析做进一步地分析研究。经过去重得到关键词共1456个,出现4次以上的关键词如表2所示。

去除"digital library",前10位的关键词分别为:metadata、ontology、information retrieval、user interfaces、user studies、architecture、web services、e-learning、education、evaluation。与数字图书馆研究领域的全景扫描中排在前10位的关键词相比较,元数据和信息检索的研究仍然是数字图书馆领域的热点,用户界面研究和用户研究、网络服务、数字图书馆体系结构、e-learning和数字图书馆教育的研究则成为近两年来研究的热点,其次,对本体、网络服务的研究进一步深化。

本文选取出现次数大于等于3的69个关键词构建共现矩阵,并进行可视化,得到的结果如图5所示。

图5 2006～2008年数字图书馆领域的关键词共现网络

2006年以来,对数字图书馆的研究内容更加丰富,研究主题间的交叉、融合的趋势比较明显。这表明数字图书馆研究的内容不断深化,涉及到的主题更加全面。传统的数字图书馆研究领域的研究内容也发生了变化。

在数字图书馆资源组织标准的研究中,元数据的研究依然是重点,但更加关注对多媒体内容的组织和描述、元数据的互操作研究和元数据收割协议OAI-

PHI 的研究。同时对本体以及与之密切相关的语义网的研究亦是数字图书馆的热点。

表 2　部分关键词及出现次数列表

关键词	出现次数	关键词	出现次数
digital library	135	annotation	4
metadata	15	children	4
ontology	10	computational science	4
information retrieval	9	digital video library	4
user interfaces	9	document clustering	4
user studies	8	FPGA	4
architecture	7	grid	4
web services	7	Human - Computer Interaction	4
e - learning	6	image indexing	4
education	6	information extraction	4
evaluation	6	information services	4
learning objects	6	internet	4
libraries	6	knowledge management	4
5S	5	CMPEG - 7	4
educational digital libraries	5	OAI - PMH	4
information visualization	5	personalization	4
performance	5	repositories	4
security	5	semantic Web	4
usability	5		

数字图书馆技术层面的研究已经由关注构建数字图书馆的基础底层技术重点转向用户界面研究(这一点可以通过与之联系的关键词的变化得以说明)。与之密切相关的是人机交互、文本检索、地理信息系统等研究内容。同时信息的可视化依然是数字图书馆技术层面的热点,信息可视化更加注重用户的体验、完善用户界面,强调人机交互,突出视频资源的可视化。信息检索方面,更加关注数据挖掘技术和信息抽取技术的研究,以及网格技术对信息检索的影响,注重对科学文献的检索。除此之外,对数字图书馆技术安全、访问控制的关注不断增加。数字图书馆技术的研究范围较广泛,从数字信息的加工到信息的利用,研究者从不同的视角切入数字图书馆技术的研究[5]。

在数字图书馆的研究中,近两年来一个值得关注的热点和研究趋向是对学习对象(learning objects)的研究,其中还包括与之密切相关的学习对象仓储建设、社区数字图书馆建设。另外一个值得关注的趋向是对数字图书馆整体模型和体系结构的研究。对数字图书馆的建设不仅仅从微观角度研究数字信息的组织、图书馆的技术,而且更加注重从宏观上整体地思考数字图书馆的体系结构,这其中对数字图书馆 5S 模型的研究较为突出。

随着数字图书馆研究的不断深化和成熟,数字图书馆开始作为一门课程进入课堂,促进了数字图书馆教育的研究,而在数字图书馆教育的学科归属问题上,研究者更愿意将其列入图书馆和情报学的学科范畴。

数字图书馆评价研究除了对各类数字图书馆本身及其可用性的评价,还包括了对数字图书馆教育和课程的评价、对数字资源标注的评价等内容。

尽管对数字图书馆服务的研究不是一个新的话题,近年来却受到越来越多研究者的重视。在数字图书馆的网络服务中更加注重为用户提供个性化的推荐服务。

3　结束语

从 20 世纪 90 年代以来数字图书馆的研究主要集中在两个大的主题:数字图书馆信息组织标准的研究和数字图书馆技术研究。这两部分主题的研究内容也经历了不断变化。目前,在信息组织和描述标准方面,元数据和本体依然是研究的热点,而在技术方面,除了信息检索这个贯穿数字图书馆领域的研究内容,用户界面和用户研究成为近两年来的热点。除此之外,对学习对象和数字仓储的研究、数字图书馆教育和评价的研究也成为近两年来研究者关注的热点。目前数字图书馆领域的研究内容更加丰富,研究层面不断深化。

参 考 文 献

1　晏尔伽,朱庆华. 1996 —2005 年 SCI - E 数据库中数字图书馆研究文献定量分析[J]. 情报科学,2007(12):1823 - 1828

2　周静怡,孙　坦. 基于 Web of Science 的数字图书馆研究论文定量分析[J]. 情报科学,2005(10):1521 - 1525

3　赵秀君. 十年来我国数字图书馆研究统计分析[J]. 图书情报工作,2005(8):99 - 102

4　钟云志,周东晓,杜香莉. 基于洛特卡定律对我国数字图书馆的研究[J]. 情报杂志,2006(6):113 - 114

5　晏尔伽,朱庆华. 数字图书馆研究进展 ——基于 LISA 数据库的文献计量分析[J]. 图书情报知识,2007(5):60 - 64

6　冯　璐,冷伏海. 共词分析方法理论进展[J]. 中国图书馆学报,2006(2):88 - 92

7　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

(责编:贺晓利)

(上接第 23 页)

15　李哲汇. 认知科学视域中的"知识管理功能"阐释[J]. 图书情报知识,2007(7):89 - 92

16　Cognitive Computation[EB]. [2008 - 11 - 20]. http://www.cog.brown. edu/Research/ErsatzBrainGroup/resources/CognitiveComputation.ppt

17　朱宝荣. 计算机模拟的认知功能及其可能性[J]. 自然辩证法研究,2003(12):72 - 76

18　郭　军. 基于分布式人工智能的知识组织[J]. 情报杂志,2004(10):33 - 35

(责编:王平军)