

基于 CSSCI 的我国“数字图书馆”研究热点 可视化分析

杜文龙(西北大学公共管理学院)

【摘要】 论文以中文社会科学引文索引(CSSCI)中以“数字图书馆”为研究主题的文献为研究对象,运用知识可视化图谱软件 CiteSpace II 对这些文献进行关键词共词分析,以可视化图谱的形式展示了 2001—2010 年我国“数字图书馆”领域的研究热点、研究前沿及其演进情况。

【关键词】 数字图书馆;知识图谱;可视化分析;CiteSpace II

【Abstracts】 Based on the papers from CSSCI with the theme of "digital library" and with the help of CiteSpace II, this paper accomplishes the co-word analysis and the hot points analysis of the objects, then demonstrates the hot points and the developing trend of "digital library" in library and information science area with knowledge map.

【Keywords】 Digital library; Knowledge map; Visualization; CiteSpace II

1 引言

随着图书馆学与信息科学的相互交融,可视化技术越来越多地应用于图书馆学的科学研究中,可视化图谱作为科学计量学代表之一,逐渐成为人文社会科学研究中的热点,利用可视化图谱,人们能够以更直观的方式看到隐藏在庞大的科技文献数据库中的各个知识单元之间的种种关联。

2 研究方法与数据来源

2.1 科学知识可视化图谱(Knowledge Mapping)

本文所采用的知识图谱可视化分析工具为 CiteSpace I-I。CiteSpace II 是由美国 Drexel 大学信息科学与技术学院教授陈超美博士研制开发的。CiteSpace II 属于多元、分时、动态的第二代信息可视化技术,可以实现自动聚类与术语标识,通过聚类视图和时区视图可以清晰地绘制出一定时期内某科学知识领域的发展趋势与前沿热点^[1]。

2.2 共词分析

共词(Co-word)分析方法的最早详细描述,是由法国文献计量学家在 20 世纪 70 年代中后期开始的^[2]。其原理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^[3]。两

个关键词共同出现在同一篇文献中的次数越多,表明二者之间的相关度越高,相似度越大,而“距离”也就越近,进而利用现代的多元统计技术,按照这种“相关”和“相似”将一个学科内重要文献的关键词加以分类,从而可以直观地显示该学科当前的研究热点^[4]。

2.3 数据来源

本文采用的数据来源于 CSSCI,数据检索策略为:时段选择为 2001—2010 年(考虑到 2011 年数据未被收录齐全),选择 CSSCI 中的来源文献;为了使采集的数据更为齐全,在所有字段检索框中输入“数字图书馆”;学科限定为“图书、情报与文献学”。检索命中 2828 篇文献。由于从 CSSCI 下载的题录和引文数据的编码类型为“ANSI”,因此需要将编码类型修改为“UTF-8”以便 CiteSpace II 能够识别并分析。

3 数据分析与结果

3.1 “数字图书馆”研究热点分析

3.1.1 “数字图书馆”研究热点术语与主题词可视化图谱表达

图 1 是 2001 年至 2010 年“数字图书馆”领域从引文中抽取出的高频热点术语和主题词共现的时区图谱。本文的热点术语来源为名词短语,图中每个三角形节点代表一个名词短语,年轮节点代表主题词,节点的大小与其被引频次成正比,节点越大,说明被引频次越多,越受到关注。图 2 是依据

中心性绘制的热点术语和主题词聚类视图,图中节点之间的连线的次数和距离的远近反映了节点的共被引频次和研究关系的疏离。

馆门户、知识管理、开源软件、数据挖掘等。

3.1.2 “数字图书馆”研究热点演变分析

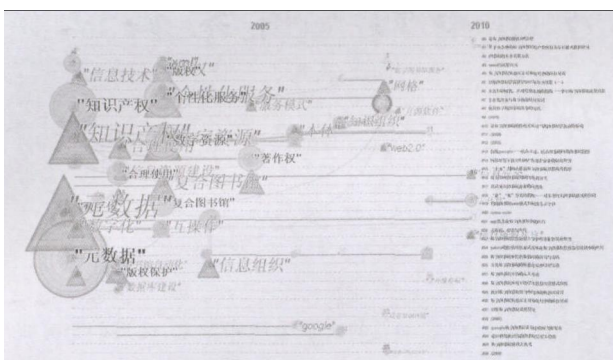


图1 “数字图书馆”研究热点术语与主题词时间分布图

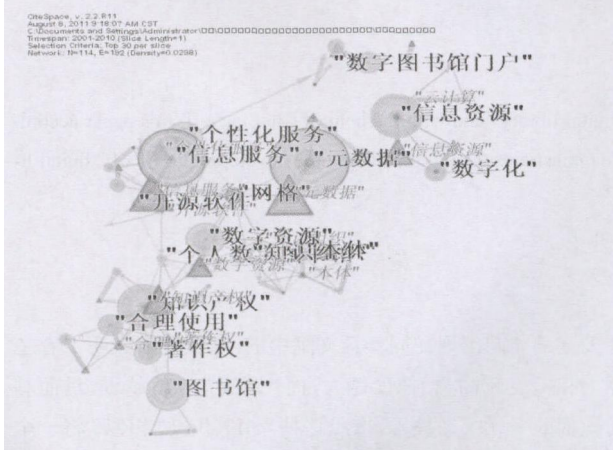


图2 “数字图书馆”研究热点术语与主题词图谱

综合图1和图2,我们可以看到一些被引频次和中心性都比较高的节点。这些关键节点共同构成了2001—2010年“数字图书馆”领域中的研究热点。从图中可以看出,5个比较大的节点中元数据为最大的节点,它的被引频次和中心性都是相对较高的,元数据是2001年研究的热点之一,聚集在周围的主题节点主要有:2000年左右形成的热点“信息组织”、2003年形成的热点“数字资源”、2006年形成的热点“网络”、2003年形成的热点“互操作”、2007年形成的热点“本体论”等。而信息服务和信息资源这两个较大的节点较近,聚集在周围的主题节点主要有:个性化服务、2008年形成的热点“信息网格”、网络、个人数字图书馆、2004年形成的热点“数据挖掘”、用户服务、开源软件、服务模式、服务质量、数字化、2009年形成的“云计算”、数字图书馆联盟、数字化等。第四个较大的节点为知识产权,聚集在周围的主题节点主要有:网络环境、个性化信息服务、2007年形成的热点“开源软件”等。第五个比较大的节点为个性化服务,聚集在周围的主题节点主要有:2005年形成的热点“个人数字图书馆”、信息组织、2008年形成的热点“web2.0”、数字图书

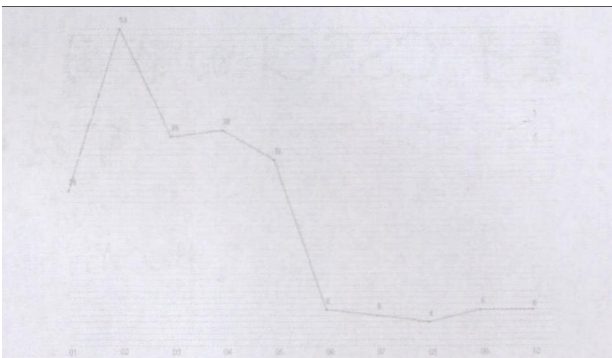


图3-1 “元数据”



图3-2 “信息服务”

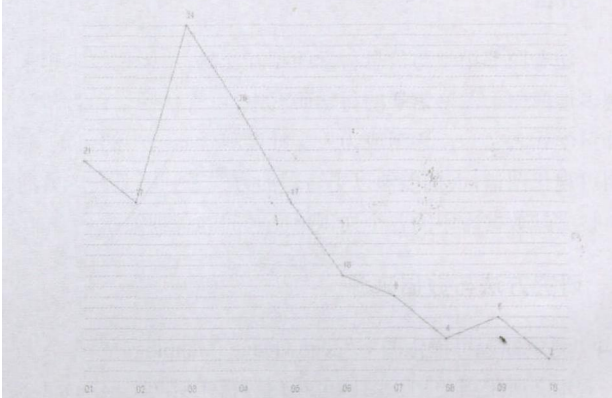


图3-3 “信息资源”

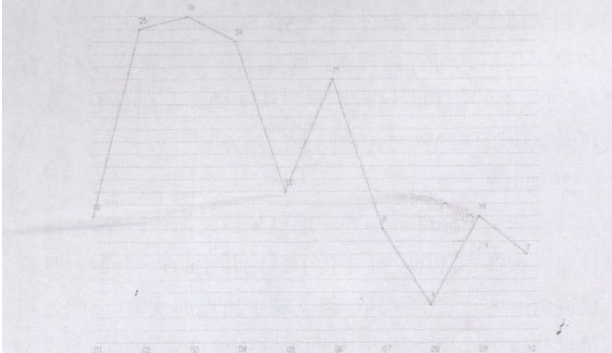


图3-4 “知识产权”

作者以“中国知网(CNKI)”收录的10家图书馆学情报学刊物为依据,对关于“知识产权”的论文进行统计,表1

和图6可以相互印证。如表1所示。

图3为“数字图书馆”领域热点主题词和术语的被引历

表1 “知识产权”研究论文年份分布

年份	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
中国图书馆学报	4	5	3	7	4	4	3	0	1	3
情报学报	2	1	2	0	0	5	2	1	0	0
图书情报工作	4	13	4	12	7	18	5	14	18	19
情报杂志	11	13	16	13	10	12	14	5	25	14
图书馆杂志	3	2	4	5	3	8	6	5	3	5
图书馆理论与实践	1	2	3	3	7	6	5	8	6	7
图书馆论坛	1	5	7	9	1	5	9	3	4	3
图书馆	3	4	4	1	2	3	6	1	4	4
图书情报知识	1	4	7	4	4	5	3	0	1	3
图书与情报	2	5	0	1	3	6	3	5	7	7
合计	32	50	50	55	41	62	56	42	69	65

史曲线图,从4条曲线中可看出这四个研究热点的演变情况。

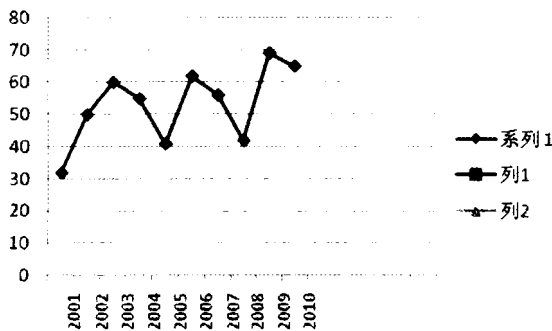


图4 2001—2010年“知识产权”历年论文统计曲线图

综合图1和图2,可以看出,“数字图书馆”研究最热、持续时间最久的主题为“元数据”,在整个网络中共现频次为208次。随着计算机技术的发展,传统目录逐渐被MARC取代,利用MARC可将文献著录的更为全面和丰富,大大地提高了检索系统的检全率和检准率。因此,在20世纪70~90年代,MARC成为最为优良的编目工具之一。但由于MARC自身的缺陷,比如结构复杂,著录项目过于烦琐等,虽保证了著录的全面性和精确性,但是著录成本高昂,对网络数字化信息反应比较迟钝,无法满足瞬间万变的网络信息资源的著录要求。

都柏林核心元素集(Dublin Core Element Set),简称为都柏林核心元数据(DC),是一个由OCLC(联机图书馆中心)和NCSA(美国超级计算机应用中心)在1995年3月联合发起,邀请来自图书馆、计算机、网络方面的学者和专

家共同研讨,由国际性使用项目Dublin Core Metadata Initiative设计的产物,目的是希望建立一套适合于网络电子资源的描述方法,以使得资源发现和检索变得更加迅速和有效^[9]。在此之后,世界各国广泛展开了元数据的理论和应用研究,并在2000年左右成为“数字图书馆”领域被普遍关注的一个热点。本体语言具有易扩展性、灵活性、知识单元呈网状结构等优点,因此在2005年之后,随着本体和本体论引入数字图书馆领域,本体和本体语言受到了该领域极大的关注,而对元数据的研究力度趋弱。

数字图书馆的起点和终点均为用户的信息需求,如何在最合适的时间以合适的方式向特定的用户提供最能满足其需求的数字服务成为数字图书馆研究的重点,因此数字图书馆总是和信息资源、用户需求、信息服务等主题词紧密结合在一起,因此不必详述此类研究热点的演变脉络。

知识产权是数字图书馆领域持续时间较长、被关注程度较高的研究热点,其被引频次为147次。美国总统信息技术咨询委员会的报告指出“知识产权的管理也将是数字图书馆面临的最为复杂和最具挑战性的问题之一。”^[10]知识产权在数字图书馆建设中具有重要的地位,数字图书馆的开发和运行是非常重视解决知识产权问题的。该领域主题词“知识产权”出现两次被引高潮,第一次为2003年,被引频次为26,第二次为2006年,被引频次为21。

从图3-4可看出,主题词“知识产权”的被引频次随时间变化趋减,但是从图4可知,在2001—2010年,以“知识产权”为研究主题的论文数量趋增,并出现两次高潮,分别

为2006年的62篇和2009年的69篇。因此可得出以下结论：“知识产权”是数字图书馆领域持续时间较长、备受关注的研究热点，这10种图书馆学刊物该主题每年的累积发文量不小于30篇；2001—2010年“知识产权”的被引频次趋减，但是该主题的论文发文量趋增。

3.2 “数字图书馆”研究前沿分析

3.2.1 “数字图书馆”研究前沿可视化图谱表达

在citespace中，研究前沿是由文献的题目、摘要、标引词和记录标识符中抽取出来的突变术语(burst terms)而确定的。这些突变术语构成某一时段某研究领域的研究前沿。Citespace在运行视图中提供了各个标引词的burst、被引频次、中心性以及被引年份等参数。

表2 “数字图书馆”研究前沿突变术语一览表(前9位)

Frequency	Burst	centrality	Keyword
48	15.7	0.36	网格
59	13.48	0.02	本体
20	11.45	0.04	云计算
28	10.97	0.00	Web2.0
32	9.03	0.00	图书馆自动化
22	9.01	0.03	开源软件
28	8.92	0.00	“google”
208	7.25	0.14	元数据
62	6.90	0.00	互操作

“google”、移动数字图书馆、数字馆藏、著作权、本体、数字图书馆门户、web2.0、网格、云计算、开源软件等等。

3.2.2 “数字图书馆”研究前沿分析

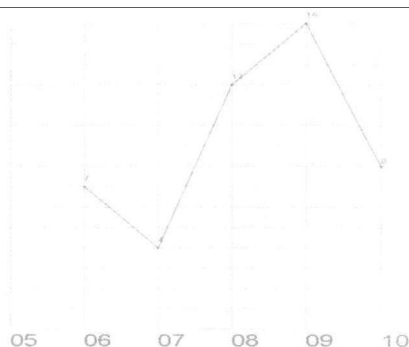


图 6-1 网格

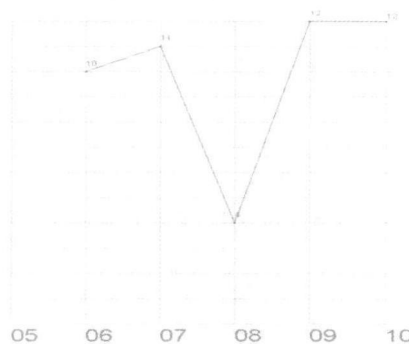


图 6-2 本体

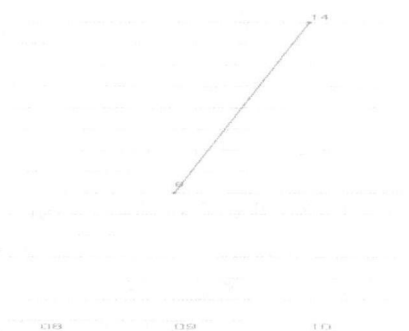


图 6-3 云计算

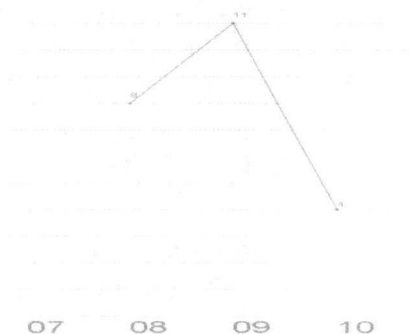


图 6-4 Web2.0



图 5 “数字图书馆”研究前沿术语时间图谱

表2是依据burst值将突现率排在前9位的标引词做降序排列。通过表2可知2001—2010年“数字图书馆”领域的研究前沿主题，其中主要包括：网格、本体、云计算、web2.0、图书馆自动化等。

图5是基于标引词和名词短语的burst值的最近几年“数字图书馆”领域研究前沿术语时间图谱，其中的红色年轮和三角形节点代表突变术语。由图5可知，近年来该领域中的研究前沿主题主要有：个性化信息服务、泛在知识环境、

图6为突变值大于10的研究前沿术语的历年被引曲线图,从图中可以看出它们的被引轨迹和发展趋势。表2中burst值排在前四位的研究前沿分别为网格、本体、云计算和web2.0,它们共同构成了2001—2010年数字图书馆领域的研究前沿,这四个研究前沿可以分为两类。

第一类为渐增型前沿主题,以“网格”为例,网格的burst值最高,为15.7,被引频次为48。由图6-1可知,主题词“本体”的被引频次逐年增加,并在2009年达到最高峰,被引频次高达15次。网格的概念和相关技术于上个世纪90年代出现。2004年,Globus联盟、IBM和HP等联合发布了新的网格标准草案WSRF,干脆把OGSI转换成了6个用于扩展Web Services的规范,发展网格被认为是下一代Internet的核心任务之一^[7]。之后世界许多国家都制定并实施了各自的网格计划。国内网格的研究起始于2005年左右。科技部通过863计划“高性能计算”的形式,在“十五”期间大力支持网格的研究和应用工作;在国家863计划及211工程支持下,我国正在建设世界最大、最先进、最实用的中国教育科研网格,并于2005年12月21日正式开通运行^[8]。这些计划和项目的实施无疑促进了国内对本体的研究以及本体技术在数字图书馆领域的应用,使得本体从2006年开始成为“数字图书馆”领域的研究前沿并成为该领域被普遍关注的研究热点。

第二类为趋弱型前沿主题,例如“web2.0”,其burst值为10.97,累计被引频次为28。数字图书馆2.0即web2.0在数字图书馆中的应用,向用户提供的是基于网络的交互式知识服务。2006年数字图书馆前沿问题高级研讨班在北京举办,刘炜做了主题为“数字图书馆2.0”的报告。之后web2.0技术被引入到数字图书馆领域,逐渐成为该领域普遍关注的一个研究热点。但是该主题词的被引频次逐年递减,2008年被引频次为8,2009年被引频次达到最高峰11,但是2010年web2.0的被引频次一度下跌至4。以“web2.0”为主题的论文在上述10种图书馆学情报学刊物中的发文量2007年为171,2008年151,2009年达到最高值217,2010年又回落至215。由此可知,图情界对“web2.0”的论文贡献数量增幅不大,此研究主题属于趋弱型前沿主题,这可以与图6-4得到相互印证。

4 结语

本文通过CiteSpace II可视化工具对“数字图书馆”领域的研究热点和动态前沿做了分析与探索。以知识图谱的方式展现了2001-2010年国内“数字图书馆”领域的研究热点

以及研究前沿的演变情况。

(1) 2001—2010年数字图书馆领域的研究热点主要为:元数据、信息服务、信息资源、知识产权、资源共享等。

(2) 2001—2010年该领域的研究前沿主要为:网格、本体、云计算、web2.0等。

(3) 近年来该领域的研究热点主要为:泛在知识环境、移动数字图书馆、数字馆藏、著作权、数字图书馆门户、网格、云计算、本体、语义网、关联数据、知识共享、概念格、开源软件等。数字图书馆前沿问题高级研讨班2009年的讨论主题为云计算、开软件、元数据、语义网、关联数据等;2010年的讨论主题为云计算、3G技术在数字图书馆中的应用、web2.0、web3.0等;2011年的讨论主题为开放获取与机构典藏、云计算及云服务、电子书阅读器、RFID、新媒体技术、移动技术的图书馆应用等。数字图书馆前沿问题高级研讨班在这三年中的讨论主题可以佐证图5中所反映的研究前沿。

总之,将信息可视化技术引入到科学研究中,通过计算机强大的计算和分析能力,形成的可视化图谱可以更直观形象的方式展现知识单元之间的各种隐性和复杂关系,有利于人们发现某一学科或主题领域的研究热点、研究前沿以及发展趋势。

参考文献

- 1 陈悦,等译.CiteSpace II:科学文献中新趋势与新动态的识别与可视化[J].情报学报,2009(3):401~421
 - 2 Law J, Bauin S, Courtial J P, Whittaker J. Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification [J]. Scientometrics, 1988 (14): 251~264
 - 3 马费成,李纲,查先进.信息资源管理[M].武汉:武汉大学出版社,2000.
 - 4 马费成,望俊成,陈金霞,等.我国数字信息资源研究的热点领域:共词分析透视[J].情报理论与实践,2007(4):438~443
 - 5 孙红玉.元数据的研究及发展[J].山东教育学院学报,2003(6):99~112
 - 6 数字图书馆:对人类知识进行普遍存取(之三)[R].中国数字图书馆工程建设联席会议工作简报,2001-04-30.
 - 7、8 李茁.数字图书馆网格[J].图书馆工作与研究,2008(1): 73~76
- (作者信息:西北大学公共管理学院硕士研究生,邮编:710127。收稿日期:2011-09-01。)

编校:刘勇定