

元数据文献共引关系的可视化分析 —利用 CiteSpace III 进行趋势分析

刘 健

(南京大学信息管理学院, 江苏南京 210093)

摘 要: 元数据是关于数据的数据, 可以用来描述数字信息资源, 促进网络信息资源的开发、管理和利用。文章采用 CiteSpace III 对元数据领域的文献共引关系进行可视化分析, 探究元数据的发展趋势。在分析关键节点的基础上, 归纳出元数据研究领域的经典文献以及该领域发展过程中出现的4个主要研究方向。这一结果有助于进一步了解元数据研究过程中的新趋势及发展脉络。

关键词: 元数据; CiteSpace III; 可视化; 共引

中图分类号: G250.7

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2013.02.016

Visualization Analysis of Co-citation Relations on the Literatures in International Metadata Field

—Research Trend Analysis based on CiteSpace III

Liu Jian

(School of Information Management, Nanjing University, Nanjing 210093)

Abstract: Metadata is the data about data. It could be used to describe digital information and improve the level of developing, organizing and using web resources. This paper provides a visualization analysis on the co-citation relation on the literatures in international metadata research fields by using CiteSpace III. Results clearly show the pivotal nodes in metadata research history. Based on these pivotal nodes, it could easily distinguish the classic literatures and four re-search fronts emerged in the metadata research history. The results will promote further understanding in emerging trends in metadata research field.

Keywords: metadata, CiteSpace III, visualization, co-citation

1 引言

文献之间的引用与被引用的关系可以反映文献与外部文献之间的关系。引文分析是根据文献间存在的相互引证的关系和特点, 对文献的引用与被引用现象进行统计分析, 可以反映一定的社会现象和情报现象, 是揭示学科发展动态的一种重要的方法。若要从总体上把握引文的特征, 分析文献的发展趋势和研究热点, 需要借助内容分析技术。文献

共引分析是一种重要的内容分析技术。共引分析是通过对特定范围内重要文献共引数据的收集, 从引文相关性特征入手, 对文献进行聚类处理, 通过对聚类形成的文献团进行分析, 进而探讨主题的发展趋势和研究热点^[1]。

CiteSpace III 是一种识别并显示研究发展趋势和研究热点的可视化Java软件^[2]。该软件对符合条件的数据进行直接分析, 不需要进行矩阵转换等, 并且能够支持多种引文数据库数据格式, 例

作者简介: 刘健(1987—), 女, 南京大学信息管理学院硕士研究生, 主要研究方向: 国内外元数据的研究进展分析。

收稿日期: 2012年10月10日。

如Derwent专利数据库、SCI数据库、CSSCI数据库等。CiteSpace III使用Kleinberg设计的跳跃检测算法检测新出现的专业术语概念^[3]作为聚类标签,标识聚类的研究热点,克服了从共引聚类中提取高频词作为聚类标签的局限,可以反映那些近期发表引用频次不高的突现词。阈值设置方法有许多种,可供研究者选择,帮助生成理想的图谱,但每种方法各有优劣。采用阈值设置方法可以对文献引文进行文献共引分析、同被引作者分析、共词分析,形成可视化共引网络,探究学科发展趋势和研究热点。为了在网络图谱中更清晰地显示关键节点,CiteSpace III采用Freeman提出的中间中心性测度检测潜在范式变化的关键点^[4]。关键节点是两个共被引文章聚类发展路径上的一系列文章,这些文章可以描绘两个研究前沿转变的特征。CiteSpace III将内容分析和聚类分析结合起来,对文献进行深层次具体微观的研究。这一可视化技术在国际上被公认为居于领先水平。

元数据是关于数据的结构化数据。从1982年Liston和Dolby提出元数据这一概念以来,其研究已经有30多年的历史。元数据最初是为了帮助描述和发现电子资源而产生的。随着研究的拓展,如今元数据描述的对象不仅仅是数字资源,也包括实体资源。例如,图书、博物馆的艺术品以及档案馆中的档案也需要使用元数据描述。国内对元数据的发展趋势的研究文献大多基于文献主题分析或对国外的专家观点的总结^[5]。例如2000—2001年,周宁和陈梅等人对都柏林核心元数据的研究重点及特点进行了总结^[6-7]。之后潘定等人对数据仓库元数据^[8]、元数据自动抽取^[9]、国外情景化元数据^[10]、社会标签^[11]等关于元数据主题分别进行了研究进展的分析。2002年梅海燕从元数据的特点、作用、可操作性、规范控制几方面进行了阶段性的总结^[12],蒋晓曦和孙坦研究了2007年国外的元数据研究进展。迄今为止,国内仍然缺乏基于文献共引网络对元数据领域发展趋势进行分析的研究。本文使用CiteSpace III对元数据的研究热点进行整体把握,有助于动态把握元数据的研究趋势,借鉴国外研究经验,促进我国元数据研究的发展。

2 调查范围、方法及调查结果

本文数据全部来源于ISI的Web of Science (SCI-E, SSCI, A&HCI, CPIC-S, CPIC-SSH)

数据库,检索策略为“标题”=“metadata”,语言类型为“English”,文献类型为“论文”和“会议论文”,时间限制为“数据库回溯最早年份—2011年”,得到原始文献1902篇,结果文献集的时间范围为1982—2011。

Citespace III参数设置是:时间段设置为1990—2011,时间分区标准为2年,由于1982—1989时间段的文献的最低引用次数和最低共引次数的统计指标基本为0,说明这段时间的文献被引量低,对共引网络的贡献不大,故不纳入统计范围。术语来源范围选择题名、摘要、系索词(descriptors,指标引文献主题的单元词或词组)和标识符,术语类型选择突现词(burst terms),使用阈值c(最低被引次数限制)、cc(各时间分区内的最低共引次数限制)和ccv(规范化的最低共引次数)限制节点数量。阈值设置为(1, 1, 0.2)、(3, 3, 0.2)、(2, 2, 0.2),如表1所示。使用pathfinder算法缩短从每个时间分区里提取网络的时间,静态显示可视化图并显示合并网络。获得761个节点,4170条连线。表1中,“引文数量”是指每一个时间分区中被引文献的数量,“节点数量”是指满足c、cc、ccv限制条件的节点数量,“连线数量”是指节点之间的共引关系连线的数量。

3 元数据研究中的关键节点

为了能清晰地看出学科发展的趋势,CiteSpace III提供时区视图以强调研究前沿和其知识基础间的顺时模式。时区视图将条目的横向变化严格限制在所属的时区内,而纵向变化则取决于同其他时区中条目的联系,帮助识别研究领域。Burton和Kebler将研究前沿对应为在短暂时间内达到被引峰值的过渡文献。研究前沿是基于新近研究成果的^[13],其引用的文献则成为知识基础,由共引聚类表示。关键节点则是连接两个研究前沿聚类路径上的文章,这些文章可以描述两个聚类发展的特征。CiteSpace III使用粗线外圈表示关键节点。关键节点可以解释研究发展的阶段特征,是指示研究方向变化的标志。图1是1990—2011年元数据研究的时区视图。图中粗线节点代表了研究方向转变的关键节点,连线代表节点之间的同被引关系。从图1中可以清楚地看到关键节点包括Sheth AP—1990, Gruber TR—1993, Lassila O—1999, Berners-Lee T—2001, Duval E—2002, Caplan—2003, Klyne G—2004以及

表1 分段网络结构表

2年时间分区	c	cc	ccv	引文数量	节点数量	连线数量
1990-1991	1	1	0.2	12	12	66
1992-1993	1	1	0.2	46	46	762
1994-1995	1	1	0.2	86	86	914
1996-1997	2	2	0.2	225	2	0
1998-1999	2	2	0.2	775	59	279
2000-2001	2	2	0.2	1458	55	216
2002-2003	3	3	0.2	2676	7	1
2004-2005	2	2	0.2	3776	91	182
2006-2007	2	2	0.2	5010	184	741
2008-2009	2	2	0.2	6252	221	821
2010-2011	2	2	0.2	3876	76	201
合计				20422	761	4170

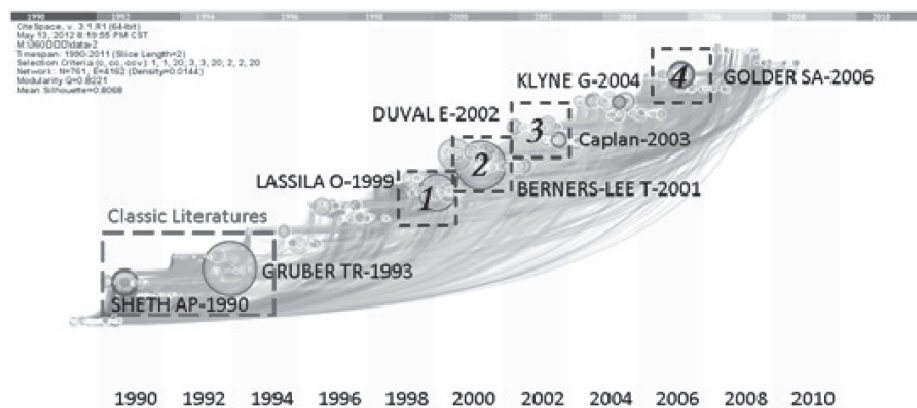


图1 元数据研究的时区视图

(外圈粗线圆环代表关键节点, 最左边的两个节点为经典文献, 右边的4个节点代表4个主要研究方向)

Golder SA-2006。这些粗线节点代表了元数据研究中出现的新趋势以及新的研究方向。

图1中最左边起始位置的两个粗线节点Gruber TR-1993和Sheth AP-1990非常突出, 从连线的方向看, Gruber TR-1993在2000以后被引用仍然活跃, Sheth AP-1990在2004年以后引用也十分活跃。这两篇文章一直处于高被引状态。

Gruber TR在1993年发表了《论可移植本体规范的转换方法》。在该论文中, Gruber TR强调了他在1992年提出的本体的定义, 即某一概念体系的规范^[13], 并在此基础上提出了一种定义可移植本体的机制。使用可移植的转化方法解决本体转化几个技术上的问题^[14]。截至2012年5月12日, 该论文SCI引用为2579次, Google引用为9574次。

1990年, Sheth AP发表了《管理分布式、异构、独立数据库的联合数据库系统研究》。在该论

文中, 作者基于系统和架构的观点定义了一个分布式数据库管理系统(A federated database system, FDBS)的参考架构, 列举了不同FDBS的开发方法。定义了一个开发普适性的FDBS架构的方法^[15]。这篇文章将联合的理念应用到数据库系统环境中, 讨论了紧耦合的FDBS的开发方法。迄今为止, 该论文的Google引用为2678次, SCI引用为40次。

研究表明, 根据引文半衰期的明显不同, 科学文献可以分为持续高被引的经典文献和在短暂时间内达到被引峰值的过渡文献^[16]。《论可移植本体规范的转换方法》和《管理分布式、异构、独立数据库的联合数据库系统研究》这两篇文章最初是为了解决异构数据库整合问题, 文中提出的可移植本体、联合的理念为解决异构元数据的互操作提供了有效的解决方法, 自从1993年发表以来持续高被

引, 成为元数据研究领域的经典文献。

4 元数据研究趋势分析

元数据研究在这20年的发展中, 各个研究方向并不是孤立存在的, 而是相互联系、相互促进的。新出现的研究方向在原来的基础上产生, 并同已有的研究方向并行发展。从图1中可以看出, 元数据研究发展过程中, 出现了4个重要的研究方向。从图1可知, 最早突现的前沿术语为元数据元素集, 以Lassila为代表。第二个突现的前沿术语为前沿术语包括信息系统和数据库系统, 以Berners-Lee为代表, 建立信息系统数据库系统是元数据应用的具体表现; 第三个突现的前沿术语为基于本体的元数据和RDF元数据, 这是元数据互操作的核心内容, 以Duval和Caplan为代表; 第四个突现的前沿术语为大数据描述、元数据存储, 以Golder为代表。

下面根据图1对元数据研究在近20年的发展历程中出现的4个主要的研究方向进行分析。

第一个研究方向的出现是以节点Lassila为代表。1998年Lassila编撰的《资源描述框架(RDF)模型和语法规则》工作草案发布^[17], 1999年《资源描述框架和语法规则》^[18]正式作为W3C推荐标准, 2004年Klyne、Lassila等修订为新版RDF标准《资源描述框架: 理念和摘要语法》^[19]。RDF是W3C规范家族中元数据的数据模型, 通常作为概念描述的通用方法或网络资源的信息模型。RDF规范自发布以来, Google引用频次达到3288次。RDF规范为研究异构网络信息资源的整合、元数据自动化管理、科学元数据方案、构建语义网提供了规范基础, 元数据规范为数据共享构建语义网作出了重要贡献。这一时期的重要的前沿术语为元数据数据集。

第二个研究方向的出现是以Berners-Lee为代表。Berners的文章《语义网》描绘了未来语义网应用的图景^[20]。截止到2012年5月13日, 这篇文章的Google被引量为13832次。他提出的“本体和元数据是语义网的核心元素”、“语义网是WWW万维网的下一次革命”、“语义网的任务”等关于语义网的观点被广泛引用。2000年Berners-Lee主持起草的W3C规范CWM——Closed world machine, 是一个语义网的通用数据处理器, 是语义网应用平台的一部分^[21]。Berners-Lee提出了语义网的观点, 代表了一个全新的发展方向。这一时期涌现的前沿术语包括信息系统、数据库系统。

第三个研究方向的出现是以Duval和Caplan为代表。2002年, Duval的《论元数据的原则和应用》系统论述了元数据与信息系统整合的方法、元数据创建和扩展原则、元数据管理、元数据应用、元数据互操作等基本问题^[22]。2003年, Caplan的《面向图书馆员的元数据基本原则》一书以论著的形式论述了元数据的基本原则, 介绍了元数据倡议和相关框架, 解释了应用配置文件的语义绑定以及增强元数据互操作的机制。关于元数据的基本概念、元数据质量、操作原则、应用等方面的系统论述是对之前工作的一个阶段性总结, 之后的研究者在进一步的研究中经常作为总结性的观点在综述部分引用, 奠定了元数据研究的理论基础。

第四个研究方向的出现是以Golder为代表。2006年在文章《协同标签的使用模式研究》中提出的协同标签的使用, 符合power-law原则、“个人标签的社会功能, 个人标签系统的推荐功能”^[23]。使用大众标签描述网络信息资源成为又一个研究热点。截至2012年12月, Golder在Google中的引用次数为1140次。大众标签是一种广泛的描述网络资源的非结构化标签, 是大量社会化标注产生的大众分类。这是一种用户驱动和群体交互的标引机制, 允许用户对资源赋予个性化标签, 并可通过标签的聚合和相关度来实现信息组织^[24]。自从Golder对大众标签研究以来, 大众标签的研究呈现蓬勃发展的态势。截至2012年12月, SCI中已有192篇文献对这一主题进行了研究。

5 结论

CiteSpace III是可视化技术的集成软件, 这一方法为研究学科领域的进展提供了一个可操作的框架。本文利用CiteSpace III软件对国外元数据领域文献的共引关系进行可视化分析, 使用软件提供的可视化的时区视图分析了元数据领域的经典文献以及近20年来的研究趋势的演化。

《论可移植本体规范的转换方法》和《管理分布式、异构、独立数据库的联合数据库系统研究》这两篇文章提出的可移植本体、联合的理念为解决异构元数据的互操作提供了有效的解决方法, 自1993年发表以来持续高被引, 成为元数据研究领域的经典文献。

元数据研究在20多年的发展过程中, 在原有研究方向的基础上, 不断有新方向出现, 基础研究与

前沿研究相互联系、相互促进,按照时间先后共出现了4个主要的研究方向。第一个研究方向以RDF规范的发布为起点。RDF是元数据的数据模型,为研究异构网络资源的整合、元数据自动化管理、科学元数据方案、构建语义网提供了规范基础。第二个研究方向以《语义网》一文的发表为起点。语义网的提出代表了一个全新的方向,W3C也制定了语义网的应用平台规范。第三个研究方向是以《论元数据的原则和应用》和《面向图书馆员的元数据基本原则》两本专著出版为起点,这两本著作系统地总结了元数据的研究成果,为之后的研究奠定了理论基础。第四个研究方向以《协同标签的使用模式研究》发表为起点。大众标签(Folksonomies)是一种广泛的描述网络资源的非结构化标签,成为重要的资源描述术语来源之一。寻找合适的词描述资源也成为这一时期新的任务。

本文利用CiteSpace III提供的时区视图清晰地展示了元数据研究在20年中发展,包括关键的节点以及研究方向的演变,为元数据领域进一步的研究提供参考。

参考文献

- [1] 黄晓鹏.我国期刊管理工作研究[M].北京:北京图书馆出版社,2003:156.
- [2] 陈超美.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,2009,28(3):401-421.
- [3] Kleinberg J. Bursty and hierarchical structure in streams[C]//8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada, 2002: 91-101.
- [4] Freeman L C. Centrality in social networks: Conceptual clarification[J]. Social Networks, 1979,1:215-239.
- [5] 蒋晓曦,孙坦.2007年国外元数据研究进展[J].图书馆建设,2009(4):107-112.
- [6] 周宁,林蓉,严正兰.都柏林核心元数据研究的新进展[J].情报科学,2000,18(6):568-570.
- [7] 陈梅.都柏林核心元数据研究最新进展[J].图书馆杂志,2001(10):31-32.
- [8] 潘定,钧毅.数据仓库元数据互操作性技术研究进展[J].情报科学,2005,23(10):1595-1599.
- [9] 曾苏,马建霞,张秀秀.元数据自动抽取研究新进展[J].现代图书情报技术,2008(4):7-11.
- [10] 李书宁,张晓林.国外情景化注意元数据的研究进展[J].现代图书情报技术,2008(2):1-6.
- [11] 翟爽,宋文.社会标签进展研究概述[J].图书情报工作,2010,54(20):41-44.
- [12] 梅海燕.元数据的研究进展[J].现代图书情报技术,2002(4):17-19,53.
- [13] Burton R E, Kebler R W. The 'Half-life' of Some Scientific and Technical Literatures[J]. American Documentation, 1960,11:18-22.
- [14] Gruber T R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing[J]. International Journal Human-Computer Studies, 1992, 43:907-928.
- [15] Gruber T R. A Translation Approach to Portable Ontology Specifications[J]. Knowledge Acquisition, 1993,5(2): 199-220.
- [16] Sheth AP. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases[J]. ACM Computing Surveys, 1990, 22(3):183-236.
- [17] Price D D. Networks of Scientific Papers[J]. Science. 1965,149:510-515.
- [18] Lassila O, Swick R. Resource Description Framework (RDF) Model and Syntax Specification[EB/OL].(1999-02-22)[2012-09-13]. <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>.
- [19] Lassila O, Swick R R. Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation[EB/OL]. (2004-02-10)[2012-09-13]. <http://www.w3.org/TR/REC-rdf-syntax/>.
- [20] Klyne G, Carroll J J. Resource Description Framework (RDF): Concepts and Abstract Syntax[EB/OL]. (2004-02-10)[2012-09-13]. <http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/>.
- [21] Berners-Lee T, Hendler J, Lassila O. The Semantic Web[J]. Scientific American, 2001,284(5):34-43.
- [22] Berners L T. CWM-Closed World Machine[EB/OL]. (2001-01-09)[2012-09-13]. <http://www.w3.org/2001/sw/wiki/CWM>.
- [23] Duval E, Hodgins W, Sutton S, et al. Metadata Principles and Practicalities[J/OL]. D-Lib Magazine. (2004-04-02)[2012-9-13]. <http://www.dlib.org/dlib/april02/weibel/04weibel.html>.
- [24] Golder SA. Usage Patterns of Collaborative Tagging Systems[J]. Journal of Information Science. 2006,32(2):198-208.
- [25] 卢小宾,孟玺,张进.基于词共现的社会化标签研究热点可视化分析[J].情报学报.2012(2):202-214.