中国图书馆学报(双月刊)2008 年第 4 期
ZHONGGUO TUSHUGUANXUEBAO

●叶 鹰

图书情报学前沿研究领域选评

摘 要 元数据与数字信息组织、本体论与知识管理技术、图书馆 2.0 与数字图书馆研究、搜索引擎与网络信息检索、智能信息处理、h 指数与学术评价六大研究领域,具有明晰的基础文献、清晰的问题域和明显的国际活力,是当前值得关注的图书情报学前沿领域。其中有的研究领域或技术方法之间存在交叉,例如元数据、语义网络与本体论相互交织、数字图书馆集成了多种技术应用等,正是当今前沿研究具有综合性和交叉性的表现,因此应鼓励面向问题的学术研究。参考文献 46。

关键词 图书馆学 情报学 前沿问题 研究综述

分类号 G250

ABSTRACT The author thinks that the following six fields are hot topics in frontiers of library and information sciences: metadata and digital information organization, ontology and knowledge management technologies, Library 2.0 and digital library researches, search engines and network information retrieval, intelligent information processing, h-index and academic evaluation. Since some fields are cross-disciplinary, the author thinks that we should encourage problem-oriented researches. 46 refs.

KEY WORDS Library science. Information science. Frontiers. Survey.

CLASS NUMBER G250

图书情报学研究是否存在一些可以客观判定的前沿领域,是一个值得讨论和探究的话题。2006 年 8 月,赖茂生教授等在中国情报学学科发展学术研讨会上做了"关于情报学前沿领域的识别与选择"的报告,用专家问卷调查法、论文统计分析法、研究项目统计分析法综合分析出 15 个领域为情报学前沿领域[1]。本文则采用客观识别与主观选择相结合的方法分析图书情报学前沿研究领域。客观识别是指通过重要基础文献、主要研究问题和国际前沿进展所体现出的历史基础和学术活力,判别 20 世纪 90 年代以来发展的新兴研究和学术热点,再经主观抉择选出了元数据与数字信息组织、本体论与知识管理技术、图书馆 2.0 与数字图书馆、搜索引擎与网络信息检索、智能信息处理、h 指数与学术评价等六大研究领域作为图书情报学前沿,并对其研究资源和研究动态进行评述。

1 元数据与数字信息组织

20 世纪 90 年代中期,伴随网络信息的增长和数字化的发展,1994 年 10 月在美国芝加哥召开的第二届万维网协会(W3C)年会期间,专家们针对互联网上的信息组织与检索问题讨论建立一套元数据(Metadata)来描述网络资源。1995 年 3 月在美国俄亥俄州都柏林市(Dublin)召开了第一次元数据会议,提出由 13 个元素(Subject, Title, Author, Publisher, OtherAgent, Date, ObjectType, Form, Identifier, Relation, Source, Language, Coverage)构成的"都柏林核心"(Dublin Core: DC),标志元数据的诞生,为信息组织尤其是数字信息和网络信息组织提供了新思想和新方法。目前承担维护和推广 DC 元数据的组织叫做"都柏林核心元数据创始"(Dublin Core Metadata Initiative: DCMI),其网站上集成了有关 DC 的大多研究资源[2]。

1.1 研究资源

第一次 DC 会议(简称 DC-1)后,很快于 1996 年又召开了 DC-2 和 DC-3,1996 年 9 月仍在都柏林召开的 DC-3 将 DC 核心集扩展成 15 个元素,奠定了被称为"都柏林核心元数据元素集"(Dublin Core Metadata Element Set: DCMES)的基础。15 个元素中内容元素有 7 个:题名(Title)、主题词(Subject)、内容描述(Description)、资源类型(Type)、来源(Source)、关系(Relation)、范围(Coverage);知识产权元素 4 个:作者或创造者(Creator)、出版者(Publisher)、其他责任者(Contributor)、权限管理(Rights);形式元素 4 个:日期(Date)、格式(Format)、资源标识(Identifier)、语言(Language)。如今,DCMES 已被澳大利亚、丹麦、英国、芬兰、美国(NISO 标准 Z39.85-2001)等国作为国家标准,被 ISO(ISO15836:2003)、IETF(Internet Engineering Task Force, RFC2413)、CEN 工作组协议 CWA(CWA/ISSS 13874-2000)等批准为国际标准,被翻译成 33 种语言。围绕 DCMES,DC 还逐步发展了一套方

法论和扩展规则,包括扩展元素、抽象模型、编码规范、应用指南、相关领域的应用纲要等,DC已不仅是简单的15个元素的集合,而是包括词表、编码规范、模型、流程、工作文件等一系列文档的标准规范体系。对此,可参考国内外专家的有关介绍[3-4]。

1.2 研究动态

早在DC-1上就确立了元数据发展的一些基本原则,这些基本原则在很大程度上影响了DC元数据的研究,为DC的发展定下了基调,主要有[4]:

简单性原则:要求定义一个能得到最广泛应用、被全球所理解和接受的最小元素集,并能作为特殊用户详细描述需求的一个核心。

易用性原则:要求能方便作者和信息提供者描述自己的文档,而不给他们增加太多的负担,并能方便地实现资源发现工具之间的互操作。

内在性原则:指DC元数据以揭示描述对象自身的内容属性为主,外部属性为辅。

可扩展性原则:希望DC作为一个核心元素集而可以通过各种方式扩展为适应各领域资源描述需要的元数据方案。

句法独立原则:指DC元数据的元素可以以多种方式编码,应用于各类技术平台中,而DC只规定元素的基本语义。

如今,已经在DC基础上发展出数十种元数据标准,通常按适用对象区分,如文本倡议TEI、档案描述EAD、艺术品著录CDWA、可视化VRA、地理空间FGDC、公用信息GILS等等[5-6]。

当前研究的关键问题有元数据互操作性等[4-8],解决元数据互操作性的一种思路是建立一个标准的资源描述框架,用这个框架来描述所有元数据格式,只要一个系统能够解析这个标准描述框架,就能解读相应的元数据格式,这就导致了XML(eXtensible Markup Language)和RDF(Resource Description Framework)的联用。

XML通过其标准的DTD/Schema定义方式,允许所有能够解读XML语句的系统辨识用XML_DTD/Schema定义的元数据格式,从而解决对不同格式的释读问题。RDF定义了由Resources、Properties和Statements等三种对象组成的基本模型,其中Resources和Properties关系类似于E-R模型,而Statements则对该关系进行具体描述。RDF通过该抽象数据模型为定义和使用元数据建立一个框架,并定义标准Schema,规定声明资源类型、声明相关属性及其语义的机制,以及定义属性与其它资源间关系的方法,还规定利用XML Namespace方法调用已有定义规范的机制。因此,XML-RDF成为元数据的重要技术支持。

元数据与数字图书馆和数字档案馆建设关系密切:在数字图书馆和数字档案馆中,元数据提供对数字资源各种属性的描述,数字图书馆和数字档案馆通过管理元数据而管理资源,元数据可以决定数字图书馆和数字档案馆的信息组织和利用方式,因此,基于元数据的信息组织[9]渐成主流,在图书馆专业方法上也有取代MARC之势。

DC传入中国后,研究者众多,以中国科学院图书馆张晓林和上海图书馆刘炜等为代表,中国科学院图书馆和上海图书馆等也分别成为国内实践DC的重要基地。

2 本体论与知识管理技术

本体论(Ontology)起源于哲学,在哲学意义上,本体是存在的本质抽象,本体论是关于存在及其本质的理论。一种本体论规定了一种哲学的元结构即第一哲学(First Philosophy),一元论、二元论、多元论等由此生成。

1993年以来,本体论的概念被大量移植到科技领域尤其是计算机相关领域,迅速兴起了工程技术形式本体论(formal ontology)研究[10]。1995年,Gruber正式提出本体设计应遵循明晰性、一致性、可扩展性、最小编码和最小承诺[11]的五条规范性原则,具有重要理论意义。2001年德国洪堡基金会在柏林将保罗奖(Wolfgang Paul Award)授予美国布法罗大学的胡塞尔哲学专家、哲学教授史密斯(Barry Smith),用于资助其将形式本体论的哲学方法和理论应用于信息科学系统,使形式本体论及其应用(领域本体论)成为引人关注的研究热点。

2.1 研究资源

形式本体论的研究资源主要涉及语义网资源和技术支持资源。语义网是万维网的发明人Tim-Burners Lee倡导的下一代万维网,旨在赋予万维网上所有资源唯一标识,并在资源之间建立起机器可处理的各类语义联系。作为本体论的前端,语义网具有桥接元数据与本体论的功能。

语义网资源以国外的WordNet(http://www.cogsci.princeton.edu/~wn/)和国内的HowNet(http://www.keenage.com)为代表。WordNet是一个基于心理

语言学的机器辞典,由普林斯顿大学的 Miller 等人研制。WordNet 2.0 包括 152059 个词(words)、115424 个同义词集(synsets)、203145 个词义(senses),描述了上下位、同义、反义、部分、整体等词汇的语义关系。HowNet 即知网,是由董振东先生逐步建立起来的。2004 版在语义方面记录中文词 75524 个、英文词 73127 个、条目总数 150100,在句法方面标引中文动词 23812 个、英文动词 19988 个、汉语句法结构 72 个。这些语义网资源可以作为构造本体论的基础。

技术支持资源包括技术标准和软件工具。目前已经出现一些核心技术标准,如资源描述框架(RDF)和 Web 本体语言(OWL)等,影响将非常深远。W3C 网站(http://www.w3.org)上还有许多相关技术资源,并可通过 RDF 和 OWL 将有关技术资源及应用联系在一起。

目前用于本体论研究开发的主要软件工具是美国斯坦福大学医学院开发的 Protégé(http://protege.stanford.edu)和德国 Ontoprise 公司开发的 OntoEdit(http://www.ontoprise.de),它们构成对本体论研究的重要技术支持,为研究提供了极大方便。Protégé 使用 Java 和 Open Source 作为操作平台,可用于编制本体论系统和知识库,可自行设置数据输入格式来输入数据,也可插入插件来扩展一些特殊功能,如提问、XML 转换等;输出格式有文本、HTML、JDBC、RDF Schema 及 XML Schema。OntoEdit 则可根据知识结构图确定类名、定义类目含义、导入专业分类表,提供本体论系统工程环境,支持构造概念、关系、定理,不依赖于某一表述语言,利用模块和插件结构,灵活性强,并能够方便地引进专门功能和词库;支持 RDF 和 DAML(DAPRPA Agent Markup Language,美国国防高级研究项目部智能代理置标语言),并能输入和输出数据库 Oracle、MSSQL、DB2 等的结构和数据;可从专业词表选词,与领域专家合作确定定义和用词,确定特性,参考元数据标准,参考专业标记语言标准,确定概念之间的相关关系等。

2.2　研究动态

目前图书情报界有关本体论的研究主要是将形式本体论作为一种新的方法论[12-15],领域本体论则沿着分类主题向语义本体"升级"的路径发展[16],直接应用是期望引导信息标引向知识标引"进化"[17]。本体论与元数据、语义网结合可应用于数字图书馆建设[18],本体学习和本体自动构建[19]则是值得关注的发展方向。

由于各知识领域对本体论的认识不同,造成一定程度的概念混乱。为此,可用三元组 T = <F,C,R> 统一形式本体论,其中 T 代表形式本体论,F 代表特定领域,C 代表元概念,R 代表元关系。一个领域的元概念建构该领域的基本术语(范畴)系统,而一个领域的元关系则建立该领域内的基本术语之间的基本联系和基本结构。若用符号"*"表示内在结合,则"元概念*元关系"也就规定了形式本体论的基本内容:元概念提供形式本体论中的基本概念系统和基本语词范畴,规定一个学科或领域的术语内涵;元关系规定元概念之间的相互关系和相互作用,使线性排列的元概念之间建立起树型或网型联络。元概念与元关系相结合,构成具有内在结构和外在功能的有机体系,这就是形式本体论三元组 T = <F,C,R>。不同领域可选不同本体,这就是可以有多种领域本体的原因;而选择反映本质特性的概念范畴作为本体才能满足 Gruber 提出的本体构建五原则。

形式本体论具有一般方法论意义,其主要价值是可以通过本体论将信息资源按照知识形式组织起来,尤其适用于网络时代的信息组织和知识组织[18]。如果说在手工操作时代要想对海量文献信息管理深化到"知识单元"层次只是一种无法实现的梦想的话,那么语义网和本体论技术的发展及其与网络技术和网格技术的结合,已经从技术手段上为实现这一梦想提供了可能。因此,语义网和本体论技术与网络技术和网格技术相结合,可能引发知识管理变革。

本体论研究的学术领域也因此归属知识管理,它将影响知识处理技术,有望成为知识组织、知识检索、知识利用的重要技术支撑。

3　图书馆 2.0 与数字图书馆研究

Web2.0 的发展和应用造就了 Lib2.0、Learning2.0、Education2.0 等,其中,图书馆 2.0 是最有特色的应用领域之一。图书馆 2.0 通常被定义为 Web2.0 的理念和技术在图书馆行业中的应用[20],和第一代数字图书馆即图书馆 1.0 相比,它在理念上强调通过将馆员主导转换为用户参与以及馆员—用户互动,在技术层面引进博客(Blog)、维基(Wiki)、简易信息聚合(RSS)等新技术,在服务层面更加关注用户体验等,从而引起了图书馆理念与实践的变革。

统观全局,真正的图书馆 2.0 范式仍未确立,国内外对图书馆 2.0 的讨论仍局限在概念的重复传播和移植上,尚未见深刻的理论推进和独创的技术发

明,因此需要继续深化研究。

目前有关图书馆2.0的研究动态[21-22]可通过资源-技术-服务轴心理解:

资源是基础。在依赖元数据和全文数字资源的同时,图书馆2.0将更多地依赖具有微结构的微内容支撑,而这些微内容则由广大网络用户参与提供,主要通过博客/播客/秀客(Blog/Podcast/Showker)、维基(Wiki)、简易信息聚合与推送(RSS/ATOM)、社会性网络(SNS)、标签与民间分类法(Tag/Folksonomy)、即时通讯(IM)等实现,图书馆2.0研究需要把资源内容建设成具有微结构的微内容,并在此基础上构造宏服务。

技术是关键。图书馆2.0是利用开放资源提供开放服务的数字图书馆阶段,其技术具有模块化、组件化特征,具有较强的平台和设备独立性,符合各类协议标准,可以方便地进行组合搭配。目前看来,Ajax(异步传输)技术与Macromedia的Flex技术、微软的Atlas技术结合而成的Ajax/Flex/Atlas技术可能成为Web2.0技术的核心内容。由于Ajax结合了Java技术、XML以及JavaScript等编程技术,使用客户端脚本与Web服务器交换数据的Web应用开发方法,打破了页面重载惯例,故能较好地优化用户体验。而对于真正实现图书馆2.0来说,则需要有核心引擎(无妨称为X引擎)来整合资源与服务,这是目前的弱点和难点。

服务是目的。图书馆2.0理念倡导个性化服务,数字图书馆也正在走向个性化图书馆时代,因此,个性化服务正是图书馆2.0和数字图书馆研究的结合点,也是数字图书馆所需要的发展模式,图书馆2.0和数字图书馆由此合流。

这样,图书馆2.0可视为数字图书馆发展的一个阶段,借助2.0的理念和技术,可望把数字图书馆研究推进到一个新的境界,这正是数字图书馆研究需要图书馆2.0的理由,也是图书馆2.0需要与数字图书馆结合的原因,而个性化数字图书馆将是它们结合的体现。图书馆2.0也只有落实到个性化数字图书馆中,才能发挥其作用。

4 搜索引擎与网络信息检索

搜索引擎(Search Engine)是以专门网站形式呈现在互联网上的检索系统,是互联网上的信息检索工具。

最早的搜索引擎是1994年问世的Yahoo!和Lycos。1995年后,搜索引擎进入了高速发展时期,演化出包括通用万维网搜索引擎(Web Search Engines)、通用元搜索引擎(Meta-Search Engines)和各种专用搜索引擎三大类型的庞大体系,被誉为仅次于门户网站(Web Portal)的互联网第二大核心技术。

搜索引擎技术发展很快,变化也快,在搜索引擎观察(Search Engine Watch)网站[23]上可以看到其动态资源。从技术上看,一个搜索引擎通常由搜索器(Robot/Spider)、索引器(Indexer/Catalog)和检索软件(Search Engine Software)三部分构成,除简单检索、加引号词组检索和逻辑组配高级检索外,搜索引擎发展的专有检索技术或方法有:

加词检索(+):在搜索词前冠以加号"+"可以限定搜索结果中必须包含+号后的词,相当于逻辑与(AND)。

减词检索(-):在搜索词前冠以减号"-"则限定搜索结果不能含有-号后的词,相当于逻辑非(NOT)。

标题检索(t:):在t:或title:后输入检索词表示仅当标题中出现同样词才符合检出要求,可使搜索结果更准确并缩小范围。

站点检索(site:):在搜索框中输入"site: DN"是在某个特定的网域或网站中进行搜索,其中DN为域名。

链接检索(link:):在搜索框中输入"link: DN"是指检索显示所有指向该网址的网页。例如,"link: www.google.com"将找出所有指向Google主页的网页。

随着自然语言理解技术的发展,有的搜索引擎如Ask(http://www.ask.com)已开始支持自然语言检索。

搜索引擎的设计希望达到快、准、全三大目标。其中快是最关键的,目前适用的搜索引擎都能在毫秒或分秒级给出响应结果,让用户等待数秒已经难以被接受;准也是关键性的,不准确的结果毫无意义;全则只是力所能及的要求,因为网络空间太大,很难求全。因此我们看到搜索引擎的使用排行不断刷新,从过去的Yahoo!变成了现在的Google[24],专业学术搜索引擎Scirus、Google Scholar也直逼DIALOG等专业检索系统"防线"。

搜索引擎是打开互联网宝库的一把钥匙,其检索技术值得关注。搜索引擎技术发展迅速,是进化最快的信息检索技术。随着互联网用户使用水平的不断提高和搜索引擎技术的不断创新,未来的搜索引擎技术的发展将更加专业化、智能化和多媒体化,而跨语

言(交叉语言)检索和跨媒体(多媒体)检索的研究与开发正成为搜索引擎技术研究的重点。

5　智能信息处理

5.1　研究资源

智能信息处理的研究资源主要涉及自然语言理解理论和实现算法。

可供参考的国外代表性自然语言理解理论(作用于英语理解)主要有:Chomsky 的转换生成语法[25-26]、Schank 的概念依存理论[27-28]、Miller 等的 WordNet 等。可供参考的国内代表性自然语言理解理论(作用于汉语理解)主要有:鲁川的句模理论[29]、黄曾阳的概念层次网络(Hierarchical Network of Concept: HNC)[30]和董振东的知网 HowNet 等。

关于自然语言理解的标准,一般按美国认知心理学家 G. M. Olson 提出的"理解四标准"判断,即已理解的话语应具备下列四属性:①可以进行变形复述;②可以正确应答问题;③可以进行摘要;④可以进行翻译。按这四条标准,机器如果理解了一种语言,就能进行自动摘要;如果理解了两种语言,就能进行机器翻译。该标准至今不能实现,表明机器不能理解自然语言。

目前常用的智能信息处理算法包括:反馈型神经网络、细胞神经网络(CNN)、遗传算法、进化算法等等[31]。智能信息处理技术一般也通过各种计算机应用算法和专门方法实现,既有通用方法,也有专用方法,前者如各种学习算法、遗传算法、进化算法、神经网络算法、BAYES 方法等;后者如用于自动分类的支持向量机(SVM)、向量空间模型(VSM)等;用于自动标引的绝对频率加权法、相对频率加权法;用于自动文摘的频率统计法、特征字串法、理解分析法等;用于智能检索的 PageRank 算法等;用于机器翻译的基于规则的方法、基于语料库的方法等。

5.2　研究动态

智能信息处理涉及的研究技术性很强,以下仅就自动分类、自动标引、智能文摘、智能检索、机器翻译等重点领域的主要方法和研究动态略作归纳。

(1) 自动分类

主要涉及两类方法:一是基于规则的方法,一般由知识库和推理机两大基础部分组成。知识库储存了从专家那里获得的关于某领域的专门知识,推理机具有推理的能力,即根据知识推导出结论,而不仅仅是简单搜索现成的答案。由于需要由知识工程师手工编制大量的推理规则,因此其开发费用是相当昂贵的。二是基于数据的机器学习方法,研究从观测样本出发,寻找规律(即利用一些做好标识的训练数据自动地构造分类器),利用这些对未来样本进行预测。现有机器学习的重要理论基础之一是统计学。传统统计学研究的是样本数目趋于无穷大时的渐近理论,现有学习方法也多是基于大数定律的结论。由于基于相对简单的机制,以及实际环境中所表现出来的良好性能,而为大部分文本自动分类系统所采用。当前研究较多的是用 KNN 作为查重算法和用 VSM 作为类特征判别算法。

(2)自动标引

主要方法有:①统计标引法:在文献中使用越频繁的实词越可能是指示主题的标引词。这种标引方法最大的优点是简单易用,而且符合人类语言应用的一般特征;②概率标引法:依据相关概率、决策概率和出现概率来决定标引词。这种标引方法目前还处于理论阶段;③句法分析法:利用计算机自动分析文本的句法结构,鉴别词在句子中的语法作用和词间句法关系 ,最终抽出可做标引词的词语。但该方法难度大效率低;④语义分析法:通过分析文本或话语的语义结构来识别文献中与主题相关的词;⑤人工智能法:让计算机模拟标引员完成标引文献的工作。

目前研究重点包括语词切分和自动标引专家系统等。

(3) 智能文摘

主要方法有:①位置法:在人工摘要中,句子为段首句的比例为 85%,为段尾句的比例为 7%,故 G. Salton 提出寻找文章中心段落作为文摘核心的思想[32];②特征字串法:文章中常常有一些特殊的线索词(短语、字串、字串链),它们对文章主题具有明显的提示作用,可以用来获取文章的主题;③频率统计法:能够指示文章主题的所谓有效词(或称实词)往往是中频和高频词;④信息提取法:常用于对一些特殊领域(如气象预报等)的文献资料做摘要;⑤框架法:借助于文章的大小标题与语义段做目次性摘要,99.8% 的科技文献标题都能基本反映主题。

目前研究重点是基于自然语言理解的智能文摘。

(4) 智能检索

智能检索与搜索引擎的发展相联系,新的智能检索技术通常首先被应用于搜索引擎,常通过运用搜索算法实现,如 VSM、PageRank 等。如 4.2 所述,目前的研究重点是跨语言检索和跨媒体检索。

(5) 机器翻译

机器翻译的基本方法可分为两大类:基于规则(Rule-based)的方法和基于语料库(Corpus-based)的方法。基于规则的方法也称理性主义方法,具有较长的发展历史;基于语料库的方法也称经验主义方法,于 20 世纪 80 年代后逐渐发展起来[33]。基于规则的机器翻译又可以分为基于转换的方法(Transform-based)和基于中间语言(Interlingua-based)的方法;而基于语料库的方法又可以分为基于统计(Statistic-based)和基于实例(Example-based)的方法。

目前研究重点是建立在双语语料库基础上的机译系统,而 WordNet 和 HowNet 则直接被作为语料库使用。

智能信息处理至今尚无通用的支持软件体系,这是一个弱项,也是一个制约发展的瓶颈。对此,研究开发 NLU 支持软件包是一个重点和难点,而基础工作首先是需要建立和完善语料库、知识库、规则库,然后在 NLU 技术上采用"语义优先、语境配合、语法辅助"原则,在系统设计上遵循"分级建设、逐步扩展、自动学习提高"的思路,才可望逐步提升到建立智能信息分析系统的境界[34]。

智能信息处理至今仍是一个边界模糊的领域,但其围绕自然语言理解形成核心的技术体系已初现端倪,一旦自然语言理解有所突破就能带动相关智能信息处理技术快速进步,而任何一个智能信息处理领域的突破都将引起图书情报研究领域的变革,故值得提早重视。

6 h 指数与学术评价

2005 年,美国科学家 Hirsch 引进一个兼顾论著数量和质量的新指标——h 指数来测评科学家的成就[35],立即引起学术界关注[36]。2006 年,Eggle 和 Rousseau 建立了 h 指数的数学模型[37],Eggle 同时提出了 g 指数[38],Batista 等提出 h_1 指数[39],使 h 指数和 h 型指数的研究[40-42]很快成为信息计量学和科学评价学的前沿研究热点。

Hirsch 将 h 指数定义为:一位作者的 h 指数等于其发表了 h 篇至少被引 h 次的论文数,即一个作者的 h 指数表明其至多有 h 篇论文被引用了至少 h 次。Braun 等将原来针对作者的 h 指数概念用于期刊[43-44],提出一种期刊的 h 指数等于该期刊发表了 h 篇每篇至少被引 h 次的论文数,或者说一种期刊的 h 指数是该期刊所发表的全部论文中最多有 h 篇论文至少被引用了 h 次。这就开始了 h 指数的推广。

一般地,设 r 是按被引次数降序排列的论文的序次,TCr 是论文 r 的被引总数,CCr 是论文 r 从 1 到 r 的累积引文数,则有以下序列:

$$r = (1, 2, \cdots r, \cdots z) \tag{1}$$

$$TC = (TC_1, TC_2, \cdots TC_r, \cdots TC_z); TC_1 \geq TC_2 \geq \cdots TC_r \geq TC_z \tag{2}$$

$$CC = (CC_1, CC_2, \cdots CC_r, \cdots CC_z); CC_1 = TC_1, CC_r = \sum_{i=1}^{r} TC_1 \tag{3}$$

h 指数和 g 指数在理论上就是:

$$h = \max\{r : r \leq TC\} \tag{4}$$

$$g^2 = \max\{r^2 : r^2 \leq CC\} \tag{5}$$

即把一位作者发表的论文按其被引次数(TC)从高到低排序(r)后,h 指数等于按被引从多到少排列的单篇论文总计被引次数(TC)大于等于 r 时对应的最大序数 r,而 g 指数则等于按被引从多到少排列的前列多篇论文累积引文数(CC)大于等于 r^2 时对应的最大序数 r。

h 指数的提出时间不足两年,系统的实证研究还比较欠缺,运用信息计量学和信息学方法,以 Thomson/ISI 编制的国际 WoS (Web of Science)、ESI (Essential Science Indicators)、JCR (Journal Citation Reports)三大数据库、Elsevier Scopus 数据库和国内 CSTPC、CSSCI 两大数据库为基础,利用 SPSS 等支持软件通过数据实验和试算发现更多 h 指数的扩展形式和类 h 指数,建立 h 指数和 h 型指数的实证支持体系,应是今后 h 指数研究的重点[45]。

最近,金碧辉、Rousseau 等在"Hirsch 核心"这一概念基础上提出了 A 指数、R 指数等 h 指数的衍生指数[46],扩大了 h 指数家族。Rousseau 等将 A 指数、R 指数和 h 指数、g 指数的关系总结为:

$$R = \sqrt{A \cdot h} \geq \sqrt{g \cdot h} \geq h \tag{6}$$

为 h 指数和 h 型指数的统一研究增添了素材。笔者带领的课题组也正在自然科学基金项目(70773101)支持下分学科、类别从作者、期刊、大学、国家、专利权人等层面展开理论与实证研究。

作为 2005 年国际上新提出的评价指标,h 指数和 h 型指数与已有计量指数结合具有发展成为下一代核心评价参数的可能,因而对其进行系统研究具有重要学术意义和价值。这是学术评价为图书情报研究开辟的一个前沿领域。

7　小结

综上所述,20 世纪 90 年代后兴起的元数据与数字信息组织、本体论与知识管理技术、图书馆 2.0 与数字图书馆研究、搜索引擎与网络信息检索、智能信息处理、h 指数与学术评价六大研究领域具有明晰的基础文献、清晰的问题域和明显的国际活力,是值得关注的实证研究型图书情报学前沿。其中有的研究领域或技术方法之间存在交叉,如元数据、语义网与本体论相互交织,智能信息处理兼及搜索引擎,XML－RDF、WordNet 和 HowNet 有演变成通用支持技术的趋势,数字图书馆集成了多种技术应用等等,这正是当今前沿研究具有综合性和交叉性的表现,因此应鼓励面向问题的学术研究,而不应以学科或领域划定边界。此外值得指出的是:国内比较关注的某些领域,如基础理论、信息分析等,在国际上并不是清晰的前沿研究领域,但由于对学科建设和优化非常重要,故完全可以作为本国特色研究或优势研究加以发展。有的研究领域,在国内外都是研究热点,如信息素养(information literacy),但历史溯源久远且问题域过于宽泛并有标准形成,故宜归属常规研究而非前沿;还有一些研究,如网络计量学(webometrics),是 1997 年后兴起的前沿研究,但问题域又过于单一狭窄,暂不独立成域。本文对不属实证研究的讨论话题则一概略之。

参考文献:

[1] 赖茂生. 关于情报学前沿领域的识别与选择[C/CD]. 中国情报学学科发展学术研讨会会议资料光盘,2006.
[2] Dublin Core Metadata Initiative[OL]. http://dublincore.org.
[3] Caplan, P. International Metadata Initiatives: Lessons in Bibliographic Control[OL]. http://lcweb.loc.gov/catdir/bibcontrol/caplan_paper.html.
[4] 刘炜等. DC 元数据的历史、现状及未来[OL]. http://eprints.rclis.org/archive/00003408/01/DCMI4 年刊_DC.pdf.
[5] 中文元数据标准研究项目组. 国外元数据标准比较研究报告[R]. 2000.12.
[6] 肖珑等. 中文元数据标准框架及其应用[J]. 大学图书馆学报, 2001, 19(5):29－35.
[7] 吴建中等. DC 元数据[M]. 上海:上海科学技术文献出版社, 2000.
[8] 张晓林. 元数据研究与应用[M]. 北京:北京图书馆出版社, 2002.
[9] 叶鹰, 金更达. 基于元数据的信息组织与基于本体论的知识组织[J]. 大学图书馆学报, 2004, 22(4):43－47.
[10] Guarino, N. Formal Ontology, Conceptual Analysis and Knowledge Representation[J]. International Journal of Human and Computer Studies, 1995, 43(5/6): 625－640.
[11] Gruber, T. R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing[J]. International Journal of Human-Computer Studies, 1995, 43(5/6): 907－928.
[12] Motta, E. et al. Ontology-driven document enrichment: principles, tools and applications[J]. International Journal of Human-Computer Studies, 2000, 52(6): 1071－1109.
[13] Poli, R. Ontological methodology[J]. International Journal of Human-Computer Studies, 2002, 56(6): 639－664.
[14] 邓志鸿等. Ontology 研究综述[J]. 北京大学学报(自然科学版), 2002, 38(5):730－738.
[15] 叶鹰. 信息科技的形式本体论研究[J]. 情报学报, 2003, 22(5): 561－564.
[16] 秦健. 实用分类系统与语义网:发展现状和研究课题[J]. 现代图书情报技术, 2004(1): 16－23.
[17] 王惠临等. 知识服务的关键技术(专题)[J]. 图书情报工作,2006(9): 6－25.
[18] 刘柏嵩. ODL:一种基于本体的新型数字图书馆[J]. 大学图书馆学报, 2005,(3): 11－15.
[19] 刘柏嵩. 面向数字图书馆的本体自动构建[J]. 中国图书馆学报, 2006, 32(5): 47－51.
[20] Miller, P. Web 2.0: Building the New Library[J/OL]. Ariadne, 2005(45): Open access at http://www.ariadne.ac.uk/issue45/miller/.
[21] 刘炜, 葛秋妍. Web2.0 技术图书馆应用分析[J/OL]. Open access at http://www.libnet.sh/sztsg/fulltext/reports/2006/libraryTech20.pdf.
[22] 刘炜,葛秋妍. 从 Web2.0 到图书馆 2.0:服务因用户而变[J]. 现代图书情报技术, 2006(9): 8－12, 67.
[23] Search Engine Watch[OL]. http://searchenginewatch.com.
[24] Brin, S. and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine[OL]. http://infolab.stanford.edu/pub/papers/google.pdf.
[25] Chomsky, N. Aspects of the Theory of Syntax[M]. Cambridge. MA: MIT Press, 1965.
[26] Chomsky, N. The Logical Structure of Linguistics Theory[M]. New York: Plenum Press, 1975.
[27] Schank, R. C. and K. M. Colby (eds.). Computer Models of Thought and Language[M]. San Francisco, CA: W. H. Freeman and company, 1973.
[28] Schank R. C. The Concept Analysis of Natural Language[M]. Natural Language Processing (Edited by R. Rustin). New York: Algorithm Press, 1973.

“国家珍贵古籍特展”在国家图书馆古籍馆举行

为配合国务院首批国家珍贵古籍名录及全国古籍重点保护单位的公布,6月14日至7月20日,由文化部主办、部际联席会议单位参与、国家古籍保护中心(国家图书馆)承办的“国家珍贵古籍特展”在国家图书馆古籍馆展出。这是建国以来规模最大、范围最广、展品最精的一次大型古籍展览。

全国80家单位和个人参展,其中既有图书馆、博物馆,也有出版社、书店等出版经营单位,还包括一些个人收藏家,涉及文化部、教育部、国家民委、新闻出版总署、宗教局、文物局、国家中医药管理局、科学院、社会科学院和军队等各个系统。

此次展出的近400件展品都是从荣登2008年3月1日国务院正式批准颁布的首批《国家珍贵古籍名录》的2392种古籍中遴选出来的,或为宋元旧本、明清精椠,或为旧钞名校、珍秘未传之本,更有生动优美的六朝隋唐写本、墨气逼人的宋明拓本、版画等,均为国家一、二级古籍。其中最早的一件为旅顺博物馆收藏的西晋元康六年(296)写本《诸佛要集经》残卷。值得一提的是,这些展品中有不少是首次公开登记、公开展出的珍贵古籍乃至孤本,如目前存世的最早、保存最好的《十三经注疏》版本等。另外,这次展览对中国多元文化也有所反映,展品除汉文古籍外,还包括33种少数民族语文古籍。

为了更好地展示中国灿烂的文化,展览从书籍史的角度划分单元,系统地向公众介绍了各个时期的官刻本、坊刻本、套印本、活字本、稿本、碑帖等古籍知识。

此次展览是全国古籍精品和古籍保护成果的一次重要展示,对于推动全国古籍保护工作,普及古籍保护知识,培养和提高公众的古籍保护意识,弘扬爱国主义精神具有重要的意义。

国家古籍保护中心

[29] 鲁川等. 现代汉语基本句模[J]. 世界汉语教学, 2000(4):11-24.
[30] 黄曾阳. HNC概念层次网络理论[M]. 北京:清华大学出版社, 1998.
[31] 王耀南. 智能信息处理技术[M]. 北京:高等教育出版社,2005.
[32] Salton, G. et al. Automatic Text Decomposition and Structuring[J]. Information Processing & Management, 1996, 32(2): 127-138.
[33] 赵铁军等. 机器翻译原理[M]. 哈尔滨:哈尔滨工业大学出版社, 2000.
[34] 叶鹰. 智能信息分析的理论基础与技术模型[J]. 情报学报, 2005, 24(2):233-236.
[35] Hirsch, J. E. An index to quantify an individual's scientific research output[J/OL]. Proceedings of the National Academy of Sciences of the USA, 2005, 102(46): 16569-16572. Open Access from http://www.pnas.org/cgi/reprint/102/46/16569.
[36] Ball, P. Index to aims for fair ranking of scientists[J]. Nature, 2005, 436(7053): 900.
[37] Egghe, L., Rousseau, R. An informetric model for the Hirsch-index[J]. Scientometrics, 2006, 69(1): 121-129.
[38] Egghe, L. Theory and practice of the g-index[J]. Scientometrics, 2006, 69(1): 131-152.
[39] Batista, P. D. et al. Is it possible to compare researchers with different scientific interests? [J] Scientometrics, 2006, 68(1): 179-189.
[40] Glänzel, W. On the h-index-A mathematical approach to a new measure of publication activity and citation impact [J]. Scientometrics, 2006, 67(2): 315-321.
[41] Liang, L. M. h-index sequence and h-index matrix: Constructions and applications [J]. Scientometrics, 2006, 69 (1): 153-159.
[42] van Raan, A. F. J. Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups[J]. Scientometrics, 2006, 67(3): 491-502.
[43] Braun, T. et al. A Hirsch-type index for journals[J]. Scientometrics, 2006, 69(1): 169-173.
[44] Saad, G. Exploring the h-index at the author and journal levels using bibliometric data of productive consumer scholars and business-related journals respectively[J]. Scientometrics, 2006, 69(1): 117-120.
[45] 叶鹰. h指数和类h指数的机理分析与实证研究导引[J]. 大学图书馆学报,2007, 25(5):2-5.
[46] Jin, B. et al. The R-and AR-indices: complementing the h-index[J]. Chinese Science Bulletin, 2007, 52(6): 855-863.

叶 鹰 浙江大学教授、博士。通讯地址:浙江大学信息资源管理系。邮编310027。

(收稿日期:2007-08-30)