综合科技信息组织的理念与实现方法研究

宋　文　孙　坦

摘　要　综合科技信息是关于科研活动主体、科学研究领域、科研过程、科研条件、科研成果、科研消息和新闻等贯穿科研过程、与科研活动有关的一切事物的信息，它强调科研活动中科技信息的多样性和复杂性。综合科技信息组织是按一定的方法和技术体系，对科研过程中涉及的科研事物进行数字化表征和有序组织，是利用先进的信息技术对综合科技信息进行集成组织与揭示，是图书馆从文献信息服务迈向知识化个性化服务的前提和基础。按照综合科技信息系统建设的目标、机制和技术体系，提出综合科技信息系统概念框架。图1。参考文献7。

关键词　综合科技信息　知识服务　本体　Web服务　信息采集　信息技术

分类号　G254　TP311

ABSTRACT　First, the author puts forward the concept of synthetic science and technology information. A discussion is made on its meaning and relation with knowledge service. Then, the author introduces some technologies to implement synthetic science and technology information organization. Finally, the author provides a frame of synthetic science and technology information system. 1 fig. 7 refs.

KEY WORDS　Synthetic science and technology information.　Knowledge service.　Ontology.　Web service.　Information acquisition.　Information technology.

CLASS NUMBER　G254　TP311

1　综合科技信息的概念和内涵

理解综合科技信息概念的内涵，首先要从信息的定义和内涵入手。经典信息论的奠基人香农(C. E. Shannon)在其《通信的数学理论》[1]中首次提出了信息量的概念。香农指出信息是用来消除随机不确定性的东西，信息量随着不确定性的消除而增大。信息概念提出五十多年来，信息科学、计算机科学、生物学、医学、哲学等领域的学者从各自学科的角度对信息的概念有过大量的定义和讨论，定义不下百种。在林林总总的关于信息的定义中，笔者比较认同邬煜和洪昆的定义。邬煜对信息的定义是：信息是标志间接存在的哲学范畴，它是物质(直接存在)存在方式和状态的自身显示[2]。该定义指出了信息与信息表述的对象(物质)的相对独立性。洪昆对信息的定义比邬煜更为详细：信息是事物及现象的存在方式之一，它是通过一定的媒介对事物及运动状态的一种显示(映射、反映)，它标志事物及现象的间接存在[3]。

结合上面两个定义，笔者认为：信息是一种间接存在，是事物存在和状态的显示；事物可以是直接存在的物质，也可以是主观存在的概念、思想等；信息需要通过一定的方式表达，通过一定的介质表现出来。

从信息的概念我们引出科技信息的概念：科技信息是一切与科研活动有关的事物和活动状态的显示。与科研活动有关的事物可以是多方面的，有客观存在的事物和精神上的事物，比如“科研人员”是一个泛指的概念，是精神层面的事物，可以用某种方式表达“科研人员”这一概念。一个具体的科研人员是客观存在的事物，这个具体的人的照片，或“人”的元数据描述是物质世界中这个“人”的显示，也就是人的信息。

综合科技信息是关于科研活动主体(如人、机构)、科学研究领域(如数学科学、物理科学)、科研过程、科研条件(如设备、实验对象)、科研成果(如专利、论文)、科研消息和新闻等贯穿科

研过程、与科研活动有关的一切事物的信息,它强调科研活动中科技信息的多样性和复杂性。

综合科技信息组织是按一定的方法和技术体系,对科研过程中涉及的科研事物进行数字化表征和有序组织。利用先进的信息技术对综合科技信息进行集成组织与揭示,是图书馆从文献信息服务迈向知识化个性化服务的前提和基础。

2 综合科技信息与知识服务

我们正处在全球化、网络化、信息化的知识经济时代。国际经济合作与发展组织(OECD)在1996年发布的报告中给知识经济的定义是:"以知识为基础的经济,是以智力资源的占有、配置,以科学技术为主的知识的生产、分配和使用(消费)为最重要因素的经济。"[4]知识经济时代呼唤知识化的服务,图书馆为顺应时代的变化和需求,正在从文献信息服务向知识服务转型。知识服务有哪些特征?它与综合科技信息组织有什么关系呢?

2.1 知识服务的特点

知识服务具有如下特点:①知识服务是基于知识单元的服务。知识服务以知识内容为单元提供深层次的知识内容服务,需要深入内容层次,对知识内容进行组织、语义关联。②知识服务是集成的服务。知识服务是知识的再创造过程,需要汲取各方面的知识,经分析提炼和创新产生新的知识。知识服务很大程度上依赖于知识的集成与关联程度。③知识服务是知识增值服务,具有创新性。知识服务是服务提供者(人或系统)通过自己独特的理解、推理和整合,在原有知识基础上产生新的知识的过程,是知识的增值服务。而传统文献服务是简单地提供原始文献的服务,没有知识的再创造过程,所以知识服务是一种具有创新性的服务。④知识服务是专业化、个性化的服务。知识服务是一项具有高度专指性和专业性的服务,需要深入到专业领域提供服务,甚至深入到课题组、科研人员中,提供个性化的服务。⑤知识服务是面向问题、提供解决方案的服务。知识服务是为解决用户的问题为目标的服务,它贯穿于用户的知识发现、知识获取、知识创造的全过程,并根据用户要求动态地、持续地组织服务。是面向用户问题,提供解决方案的服务。⑥是需要依靠先进技术的服务。知识服务需要依靠大规模的数据仓库技术、信息集成和互操作技术、数据挖掘技术、知识组织技术、智能推理技术、网格技术等等,依靠先进的技术提供集成化的、精深的知识服务。

2.2 综合科技信息组织与知识服务的关系

综合科技信息组织与知识服务之间有着密切的关系,表现在以下方面:

首先,综合科技信息是关于科研人员、科研机构、科学数据、科研仪器设备、种质资源、科研新闻、科技成果、科技政策等方面的信息,综合科技信息组织就是对这些科技资源进行表述和深度组织,是以知识对象为单元的组织与揭示,可以为知识服务提供以知识内容为单元的信息支持。

其次,综合科技信息组织对各类科研信息进行集成组织,对各类对象之间的内在关系进行语义关联,比如科研项目与科研人员,他们之间的关联关系是科研人员主持并参与科研项目。知识化服务需要这样的内容集成组织与揭示体系。

第三,综合科技信息组织需要按本体的方法组织各种内容,按照本体的方法构建内容对象的语义关系和添加公理后,使综合科技信息系统具有了自动的分析和推理能力,用户可以提出具有语义的问题,系统可以根据用户的提问进行一定程度的逻辑推理,而不仅仅是简单的词语匹配,可以提供面向用户问题的服务,为用户提供问题解答,所以综合科技信息的关联组织与知识服务有一定的相似性。

第四,综合科技信息的组织、综合科技信息服务系统的建设需要对当前先进的网络技术、语义技术、信息自动采集与标引技术进行研究和应用,综合科技信息系统依赖的技术体系与知识服务依赖的技术体系具有共同性。

3 综合科技信息组织的方法与技术

综合科技信息的组织与揭示工作,其本质的目标和意义是采用语义技术、信息自动采集和注释技术、数据挖掘技术等国际前沿技术,采集和集成组织综合科技信息资源,使综合科技信息系统成为科研资源集成揭示平台、科研交流与合作平台、科技成果展示平台、科技信息聚合查询和获取平台。综合科技信息系统将为知识化服务提供基础保障,最终成为 e-Science 的有机组成部分。

综合科技信息从采集到组织关联再到最终的应用与服务,整个过程需要应用的技术方法包括几个方面:

3.1 本体技术

本体是下一代互联网 - 语义网的核心技术。斯坦福大学的 Gruber 在 1993 年首先引入本体概念,定义为"An explicit specification of a conceptualization"。Borst 在 1997 年对该定义做了补充和修改,修改后的本体定义是:共同认可的概念体系的明确的、形式化的规范说明(an explicit formal specification of a shared conceptualization)。本体由概念(concept)、关系(relations)、属性(property)、公理(axioms)和实例(instances)五个元素组成。

概念:表示具有一个或多个共同特征的个体的集合。

关系:是概念之间的关系,有等级关系和其他特别关系,等级关系构成了本体概念的树状等级体系,其他关系定义了一个概念与另一个概念的特定关系,使本体成为一个网状结构的概念体系。

属性:用于建立两个个体之间的关系,属性有定义域和值域。定义域是定义该属性的类,值域是实例属性值的个体所属的类。

公理:是概念、关系、属性之间的一些约束条件。

实例:概念的实例是一个个体,其特性符合概念的内涵标准。

本体需要采用形式化语言进行描述,以便可以被机器所理解。W3C 推荐的本体描述语言 OWL 专门用于描述本体,OWL 语言的语法基础是 XML。

用本体的方法来组织综合科技信息,基本方法是对综合科技信息进行分类和归类,提炼出代表各种科技事物和对象的概念,建立概念等级和概念关系。用这个概念体系为基础对各种科技信息进行组织,赋予实例以属性,并根据需要添加公理,建成具有丰富语义关系的综合科技信息知识库。

本体的建设和应用是一个复杂的课题,也是目前国际语义网研究领域的热点问题。本体建设中需要研究的问题包括本体的存储与查询技术、本体工程学,本体的集成、合并与联合,本体的进化,本体推理、本体学习和本体应用等等。

3.2 语义门户技术

语义门户首先是一个门户,它以本体为基础,是面向特定用户群,为了进行信息交流的目的而建设的。语义门户除需要构建基础的本体外,为实现基于本体的语义服务,需要专门的语义门户技术支持。

语义门户的框架研究是目前语义门户建设中的基础问题,德国卡尔斯鲁厄大学应用信息学与形式描述方法学院(AIFB)建立了标准的门户框架 SEAL,是目前许多语义门户参照的基础框架。

语义门户需要语义浏览和语义检索技术,用户可以用语义模板编辑查询问题,语义检索按用户问题进行逻辑查询。需要研究语义相关度的算法,对检索结果按语义相关度进行排列。语义浏览提供按概念等级和概念关系浏览本体实例,需要采用可视化技术。

语义门户提供语义编辑技术,用户可以按本体关系建立内容实例。语义门户还需要建立与相关内容门户的内容联合体,通过本体联合、内容集成、虚拟集成查询等方式,扩充语义门户的知识内容为用户提供丰富的内容查询服务。

综合科技信息按本体组织成知识库,可以

采用语义门户技术,为建立综合科技信息语义门户提供服务。

3.3 Web Service 技术

Web Service 技术代表了基于万维网的分布式计算,它提供了一套技术和方法体系,使得在互联网上进行资源、服务的共享成为可能。Web Service 的技术体系包括 SOAP、WSDL 和 UDDI。SOAP 提供了一种应用程序与 Web 服务进行通讯的机制,WSDL 提供了一种向其他应用程序描述 Web 服务的方法,而 UDDI 用于创建 Web 服务注册中心,提供对 Web 服务的集中发现机制[5]。

综合科技信息采用 Web Service 技术的重要意义在于实现广域网上的综合科技信息资源与服务的集成。综合科技信息资源的建设不是一个系统或一个机构的信息内容所能涵盖的,综合科技信息体系是语义网上分布的多个同构或异构系统的集合。Web Service 提供了将分布的综合科技信息系统进行登记、提供集成服务的标准化的技术方法。采用 Web Service 技术体系,使综合科技信息服务从单一的、小范围的服务扩展到综合的、范围更广的、可以与互联网其他服务信息有效集成的服务系统。

3.4 Web 信息采集技术

万维网上存在着大量的综合科技信息,对这些信息进行采集与组织,是综合科技信息资源建设的重点任务之一。

Web 综合科技信息的采集需要采用 Web 信息采集技术。Web 信息采集是一项综合性技术,主要由两个方面的技术成分构成:

(1)网络信息采集技术。通常搜索引擎采用的是页面爬行器(Web Crawling),也常称作 Web Spider、Web Robot 或 Web Worm。主要原理是按照初始的 URL 对页面进行爬行,通过页面的超链接爬行到更多的页面。这种网络信息采集模式主要用于大规模的搜索引擎,一般不适用于面向特定领域或特定要求的信息采集。主题搜索成为近几年的研究热点,国内外在主题搜索的算法上有很多研究,创建了一些经典的方法,如 S. Chakrabarti 开发了一个典型的基于主题的 Web 信息采集器[6],它的主题集是用样本文件来描述的。采集系统首先保存一个经典的主题分类(例如 Yahoo 的主题分类),并且为每一个主题分类都保存若干个内容样本,用于详细地刻画这一类主题。微软中国亚洲研究院开发了面向对象的垂直搜索系统,可以进行对象(如产品)搜索[7]。主题搜索和面向对象的垂直搜索将成为下一代搜索引擎的主流。

(2)信息抽取技术。信息抽取技术处理网络上大量非结构化或半结构化的信息资源,对其进行分析,抽取对象元数据,识别对象之间内在关系,按预先定义的语义关系建立对象之间关系。信息抽取技术也是当前信息技术领域的热点研究课题,涉及自然语言处理技术、自动标引技术和页面分析技术等。

综合科技信息的采集与组织需要采用主题搜索和对象搜索技术对网络上的综合科技信息资源进行采集,采用信息抽取技术实现自动或半自动的信息抽取与标引,结合语义门户的语义编辑功能,才能实现全面快速的综合科技信息资源的组织与建设。

4 综合科技信息系统概念框架

按照综合科技信息系统建设的目标、机制和技术体系,我们设计了综合科技信息系统概念框架。概念框架为三层结构:数据集成层、存储层和 Web 服务层(见图 1)。

4.1 数据集成层

建设包含大量丰富信息的综合科技信息系统的关键是 WWW 上大规模的异质数据的集成。网络上存在大量非结构化、半结构化的信息资源,如 HTML 格式信息、XML 格式信息、数据库格式信息等,对异质数据的处理主要是对这些数据格式进行封装,即将非结构化(如 HTML 格式)、半结构化格式(如 XML、数据库格式)提升到 RDF/OWL 格式,本体将作为异质数据输入的语义模型。

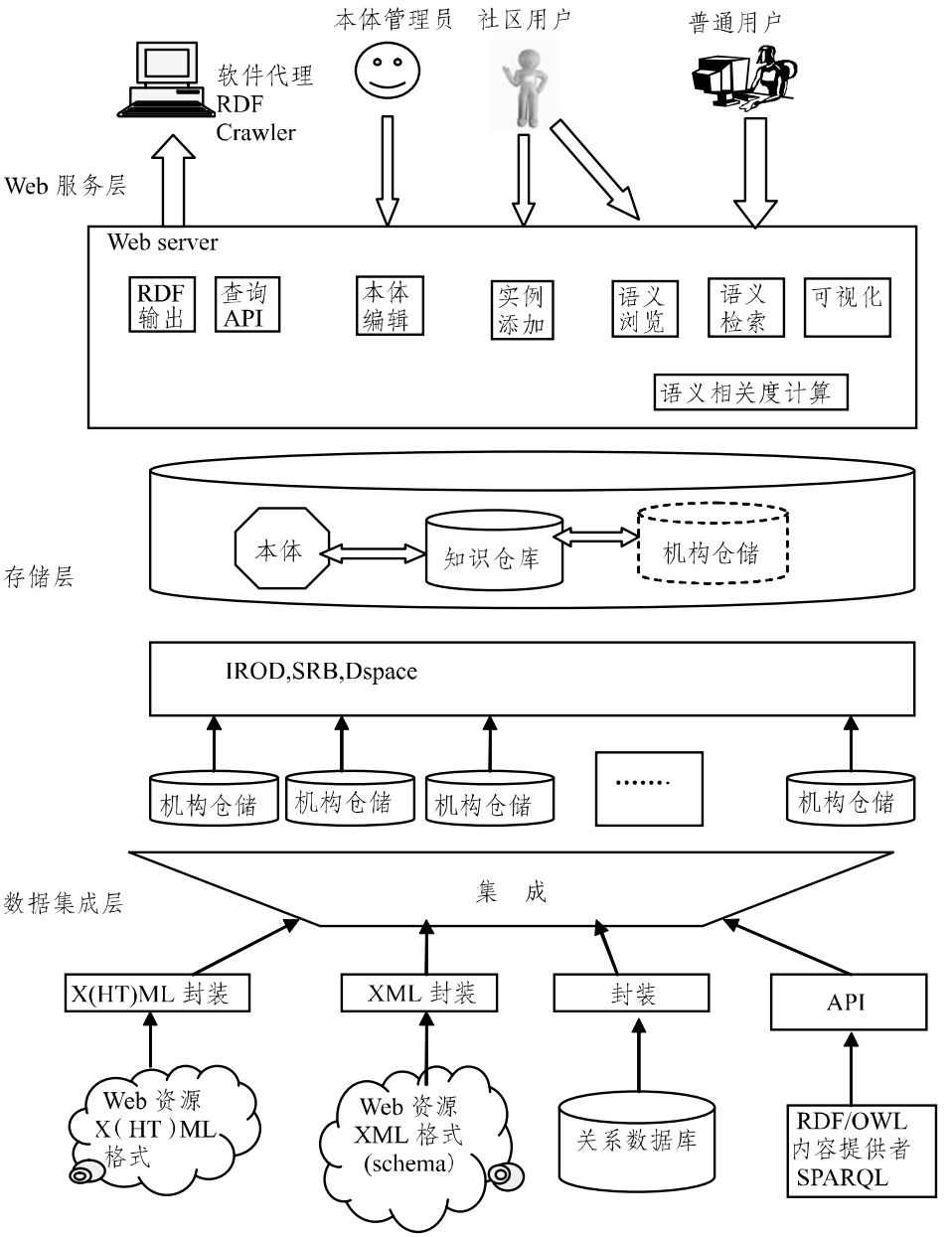

图 1 综合科技信息系统框架

4.2 存储层

存储层存储综合科技信息的元数据信息、全文信息和组织这些信息的知识关系的语义模型——本体。存储层的核心组件是本体,本体作为建立综合科技信息各种对象之间关系的知识组织模型。为达到面向各领域、各课题组、各研究所的综合科技信息资源可以动态集成和互操作,需要有一个共同的知识组织模型。

在存储层中,本体是各系统共用的,可以集中存储,也可以分散存储,但需要统一更新。知识库是对综合科技信息的元数据描述,按照本体的语义模型进行组织,各系统可以有自身专门的知识库系统机构仓储存储各种格式的全文信息,机构仓储通过知识库的元数据信息建立内容对象之间的关联。机构仓储可以采用数据网格的技术分布存储在不同的计算机上,通过网格浏览器获取内容信息。

4.3 Web 服务层

供四种类型的用户通过 Web 服务器与系统交互:

首先,系统通过 RDF 发生器提供知识库信息,远程应用(软件代理)RDF 爬行器采集存储层内容,也可以通过查询 API 直接访问存储层的本体库、知识库和机构仓储的信息内容。第二,本体管理员通过本体编辑器建设、更新本体中的类、属性和关系。第三,社区用户提供知识库和机构仓储内容。社区用户通过系统提供的基于本体的模板,提交个人信息、研究领域、研究项目等综合科技信息内容,提交论文、报告等全文信息内容,系统按本体的知识关系将社区用户的信息保存在知识库和机构仓储系统中。第四,社区用户和普通用户能够访问语义门户,访问可以有三种形式:基于本体的语义浏览、语义检索和语义可视化。检索结果经过语义相关度分析,按语义最接近顺序排列,检索结果根据属性关系关联到另外的知识或事实内容。

5 结语

从用户需求和技术环境的变化出发,中国科学院国家科学图书馆在 2007 年首次提出了综合科技信息的概念,并组织了专门人员对有关的机制、模式、政策以及相关的技术体系进行研究。这是一个全新的研究和建设领域,我们热切希望图书馆界和信息技术界同行共同努力,为用户建设一个更加丰富、方便、快捷的信息环境。

参考文献:

[1] 香农. 通信的数学理论[M]. 上海:上海市科学技术编译馆,1978.
[2] 邬焜. 信息世界的进化[M]. 西安:西北大学出版社,1994.
[3] 洪昆辉,杨娅. 论信息存在的复杂性[J]. 云南社会科学,2005(6):35-40,45.
[4] 经济合作与发展组织(OECD). 以知识为基础的经济[M]. 北京:机械工业出版社,1997.
[5] 宋文,孙坦,周静怡,等. 科技信息资源与服务集成揭示系统中分类体系的设计[J]. 图书情报工作,2007(7):71-74.
[6] S. Chakrabarti. Focused Crawling[OL]. [2007-10-14]. http://www.cse.iitb.ac.in/~soumen/focus/.
[7] Zaiqing Nie. Object-level Vertical Search[OL]. [2007-10-14]. http://research.microsoft.com/users/znie/cidr2007-nie.pdf.

宋 文 中国科学院国家科学图书馆研究馆员,硕士生导师。通讯地址:北京北四环西路 33 号。邮编 100190。

孙 坦 中国科学院国家科学图书馆研究馆员,博士生导师。通讯地址同上。

(收稿日期:2008-06-20;
最后修回日期:2008-10-05)