信息管理学术前沿文库

XINXIGUANLI XUESHU QIANYANWENKU

信息资源网络模型及应用

王 昊 著

南京大学出版社

国家社会科学基金项目“面向语义网本体的知识管理研究”（09CTQ10）

信息管理学术前沿文库

信息资源网络模型及应用

王 昊 著

南京大学出版社

图书在版编目(CIP)数据

信息资源网络模型及应用/王昊著. —南京:南京大学出版社,2010.5
(信息管理学术前沿文库)
ISBN 978-7-305-06984-0

Ⅰ. ①信… Ⅱ. ①王… Ⅲ. ①社会科学-情报检索-研究 Ⅳ. ①G252.7

中国版本图书馆 CIP 数据核字(2010)第 077386 号

出版发行 南京大学出版社
社　　址 南京市汉口路 22 号　　邮　　编 210093
网　　址 http://www.NjupCo.com
出 版 人 左　健

丛 书 名 信息管理学术前沿文库
书　　名 信息资源网络模型及应用
作　　者 王　昊
责任编辑 孙　辉

照　　排 南京大学印刷厂
印　　刷 南京大学印刷厂
开　　本 787×960　1/16　　印张 22.75　　字数 385 千
版　　次 2010 年 5 月第 1 版　2010 年 5 月第 1 次印刷
ISBN 978-7-305-06984-0
定　　价 46.00 元

发行热线 025-83594756　025-83686452
电子邮箱 Press@NjupCo.com
Sales@NjupCo.com(市场部)

《信息管理学术前沿文库》总序

随着社会信息化的迅速发展，信息管理领域所涉及到的理论、方法和技术也发生了巨大的变化，网络的广泛渗透和信息技术的日新月异，以不可思议的速度改变着人们的学习、工作和生活方式，也给信息管理学科带来严峻的挑战和前所未有的机遇。信息管理的理论需要创新、方法有待突破、技术亟需引进和改造。正是基于这样背景，我们组织了《信息管理学术前沿文库》，期望能够推出一批适应时代需求的信息管理研究新作。

长期以来，图书馆学、情报学、档案学的学科地位的讨论一直是信息管理领域的热门话题之一。虽然这个学科在其发展过程中已经产生了很多理论成果，但在把其他学科大量理论移植到本学科时，也具有了某种程度的依附性。我们认为，一门基于信息资源保障和服务于其他行业的学科，它在学科之林中的地位并非也是依附性的。换言之，独立的学科体系，应在学术界和社会上具有一定的影响力和辐射力。高等学科教育是以学科理论为基础建立系科专业的，学科建设的重要任务是建立一个有内在逻辑联系的、充实的学科理论体系，并以此为依托，生长出面向社会需要的方方面面。因此，我们要做的不再仅仅为图书情报学科的各个分支寻找适用的理论，而是将现有理论整合，将经验性的、仍处于“潜学”的知识归纳起来，将已经产生一定影响力的理论成果和技术方法进行系统地总结和阐释，上升为学科理论。这套文库的推出，正是从弘扬本学科最新研究成果，推出最新理论和方法为指导思想而策划的。

我们清楚所面临的任务的艰巨性，这对作为文库主要作者群的南京大学信息管理系的中青年教师既是一个挑战，也是一个机遇，因为完成这套文库不仅对本系的学科建设具有重要的意义，对提升教师们的学习能力和研究能力也将是一次宝贵的磨练机会。图书馆、情报与档案业界和学界的同仁们已经在这片土壤中进行了卓有成效的开垦，在基础理论、信息描述、信息获取、信息传递、信息处理与重构、信息控制、信息组织、信息利用和吸收、信息预测、信息评价等领域的理论探索方面打下了坚实的基础，成为文库作者们汲取理论养分的巨大资源。这套文库除了继续跟踪上述领域的研究，并将这些领域的前沿成果展现给读者，还可以通过向兄弟院系理论

工作者和业界专家学习，提高作者自身的教学科研水平。为此，我们诚恳邀请国内外本学科知名学者担任审稿专家，也希望读者们给予指导和帮助，通过这个平台进行交流。我们期待着这套文库在本学科的发展中发挥作用，也期望这些工作能为创建一流学科贡献绵薄之力。

《信息管理学术前沿文库》编委会

2010 年 1 月

序

本书作者王昊博士在硕士阶段就跟随我学习，毕业后又考入我的博士，2008年获得博士学位。他学习期间刻苦专研，为专业研究打下了较为坚实的理论基础和应用能力，其博士论文也写得很好，评审专家对他的博士论文评阅和答辩都给予了优秀的成绩。当他向我征询是否可以将博士论文修改后出版专著时，我积极鼓励并支持他完成这项工作。经过他一年多的努力，书稿终于完成，我应允写上数语，是为序。以鼓励他在未来的研究中取得更多更大的成就。

本书的主题是利用本体技术构建学术资源网络模型，这是一项颇具挑战的工作，它突破了传统的学术资源关联结构，把语义网技术用于构建学术资源，使学术资源发挥了更大的功效。过去我们对学术资源的分析往往仅限于文献间的联系，很难将作者、机构、主题、学科、期刊、论文、图书以及研究热点等学术资源构成多维、复杂的知识网络来综合考察学术资源，但本体技术可以做到这一点，本书正是利用这一技术构建了学术资源语义网络，并借助于这个网络进行了学术影响分析，还进行了针对学术资源的语义检索探索。可见其研究成果对相关领域有很大的参考借鉴价值。

1999年我设计了《中国人文社会科学引文索引》(Chinese Society Science Cited Index，简称CSSCI)，但限于数据库本身的功能，CSSCI很难体现其复杂的信息关联和语义表达，尤其是通过它来反映各类学术资源的有机联系较为困难，王昊博士把本体技术用于CSSCI的数据架构，并借助其进行数据分析和科学评价，可以说是我们早期工作的延续，对开发CSSCI的新功能有很大的帮助。

本书通过对本体基础理论的研究，并有效的将其用于描述领域内知识，实现领域知识的语义理解。针对CSSCI学术资源的现有数据组织方式的局限性，利用本体技术优化了原有数据结构，以面向对象形式来描述人文社会科学领域的学术资源，从而为CSSCI学术资源语义检索和挖掘潜在知识等问题提供有效的解决途径。

对于呈现在读者面前的这本专著，我认为其价值主要有这样几个方面：

第一，在总结分析国内外现有的本体构建方法和关键技术基础上，结合CSSCI信息服务现状和CSSCI本体的特点，提出了适合CSSCI本体构建的指导思想和6步骤的循环建模过程。

第二，建立了CSSCI本体概念模型。从CSSCI原始关系数据库中抽取主要概念，并通过核心扩展的方法获取辅助概念和下位概念，建立起比较完整的CSSCI概念层次结构；定义每个概念的属性，具体描述概念属性的计算方法，为学术资源本体的实例化奠定了基础；在本体概念模型建立完毕之后，可以采用Protégé工具对其进行具体描述、图形化展示以及逻辑检测。

第三，CSSCI大规模数据的语义标引。在概念模型的指导下，充分利用CSSCI原始数据结构，采用函数依赖、数理统计、TF－IDF算法、形式概念分析和机率模式算法等对大规模来源数据进行语义标引。在完成各类实例属性值设置后，CSSCI实例以面向对象的形式被集成在一个巨大的学术资源网络中。书中还对该学术资源网络的存储方式以及评价方法做了探讨。

第四，建立了基于CSSCI本体的知识检索服务平台。在提出基于本体信息检索系统一般模型的基础上，结合CSSCI知识检索服务功能和特点的分析，建立了适合CSSCI用户的知识服务平台的系统架构，并开发了一个原型系统用以验证，试图提供CSSCI学术资源的语义检索功能。

第五，提供了基于CSSCI本体的引文分析服务。本体最大的优势在于将领域知识以面向对象的结构进行组织，与对象相关的所有知识被存放在对象属性值中。因此，通过对对象属性知识的深入分析和相互比较，可以对对象产生深刻的认识。在这种思想的指导下，书中对CSSCI本体实例库进行了多对象、多方位、多维度的引文分析，借助多维尺度分析方法和数据挖掘技术，实现了CSSCI学术资源的关联分析、热点分析以及发展趋势分析，得到了一些可参考的，较为全面准确的结论。

国内目前对于本体的研究呈现出强劲的势头，然而大多是构建理论模型，缺少实证研究。作者在采用文献调研、专家咨询、案例分析、方法支持、模型构建、系统开发、实验论证以及小组讨论等研究方法的基础之上，注重理论方法与实践应用相结合，力图在促进我国本体机制研究发展方面做出一定贡献。

王昊同志作为一名青年学者，虚心好学，勤于思考，他作为优秀的博士生毕业后选留本校做教师。可以说，博士毕业既是他学业的终结，也是他事业的起点，这本书是他的第一本学术专著。作为他的导师，衷心的期望他以这本著作的出版为契机，不断进步和发展，为我国信息管理事业的发展做出更大贡献。

苏新宁

2010年1月7日

目　录

第1章 引 言

1.1 语义网和本体

网络数据的激增，宣告“信息爆炸”时代的到来。因特网成为了人们获取数据的主要信息源，人们往往需要花费大量的时间和精力浏览网页、搜索信息、筛选数据，数据检索、访问、显示、整合和维护都变得非常困难。在这种情况下，Web 的创始人 Berners-Lee 在 1998 年首次提出了语义网的概念[①]。

➢ 语义网

1998 年，Web 的创始人和 W3C(World Wide Web Consortium，全球互联网联盟)组织的执行主席 Berners-Lee 首次提出了语义网[②]的概念。在 2000 年 12 月 XML2000 大会的重要发言中，Berners-Lee 正式提出了语义网概念；2001 年 2 月，W3C 正式成立“Semantic Web Activity”来指导和推动语义网的研究和发展，语义网的地位得以正式确立；2001 年 5 月，Berners-Lee 等在《Scientific American》上发表了文章《The Semantic Web》，用浅显语言和生活实例从“应用设想”、“意义表达”、“知识表示”、“ontology”、“Agent”以及“知识演化”等诸多方面对“语义网”作了较全面的阐述。

Berners-Lee 指出“语义网”中的“语义”是指“机器可处理的”，而不是指自然语言语义或是人类推理语义。语义网是一个使用能够表达语义(或机器可处理)的元素来描述信息，以满足智能软件代理对异构、分布信息的有效访问、合理交换、语义处理和准确检索等要求的公开环境。其核心想法是创造元数据来描述数据，进而使计算机能处理它。它具有 3 个方面的特征：① 需要充分利用数据。② 可转化已备将来使用。③ 处于公开的环境中。图 1-1 展示了语义网中对象之间的语义关联。

---

① http://www.w3.org/DesignIssues/Semantic.html. [2008-02-20].

② http://www.xml.com/pub/a/2000/12/xml2000/timbl.html. [2008-02-20].

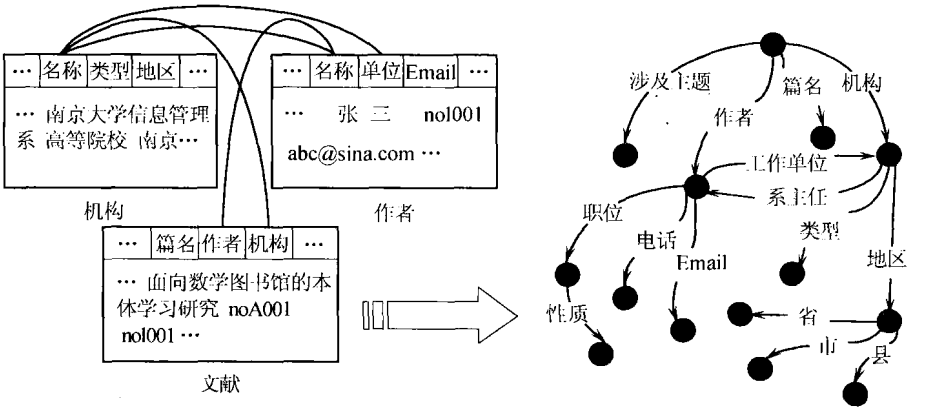

图 1-1 语义网中对象的语义关联

Berners-Lee 提出了语义网的七层体系结构[①]，认为自底向上分别为：UNICODE 和 URI、XML、RDF、Ontology、Logic、Proof、Trust，如图 1-2 所示。① Unicode 和 URI(Universal Resource Identifier)定义了编码级标准，是语义网的存储基础，URI 负责资源的标识，Unicode 负责资源的编码。② XML(eXtensible Markup Language)是半结构化的文档或描述格式，定义了语义网语法级的标准，为上层提供语法描述文档。③ RDF(Resource Description Framework)提供了一套标准的数据语义描述规范，通过谓语来描述语义，定义了最小要求模型，包括类(或子类)和属性(或子属性)以及取值范围、注释等，是语义网的数据互操作层。④ Ontology 也是数据级的描述格式，是语义网的知识集合，它定义了更多的元信息，例如及物性质等，具有唯一性、明确性、形式化和共享性等特点，Ontology 具有广泛的互用性和互变性，可以使用 OWL(Ontology Web Languages)、OIL(Ontology Inference Language)、DAML(DARPA Agent Markup Language)以及 SHOE(Simple HTML Ontology Language)等标准化本体描述语言描述。⑤ Logic 层定义了规则及其描述方法，作为自动推理的基础，它使用规则语言如 ORL(OWL Rules Language)等描述单一逻辑，为上层提供推理规则，对 Proof 进行验证，一般的，Logic 层不存在标准的推理引擎，多种引擎的推理能力不统一。⑥ Proof 层使用 Logic 层定义的推理规则进行逻辑推理，

① http://www.w3.org/2000/Talks/1206-xml2k-tbl/Overview.html. [2008-02-20].

得出某种结论。⑦ Trust 层认为在可信任的数据上进行可信任的推理，得到的最终结论也应该是可信任的。

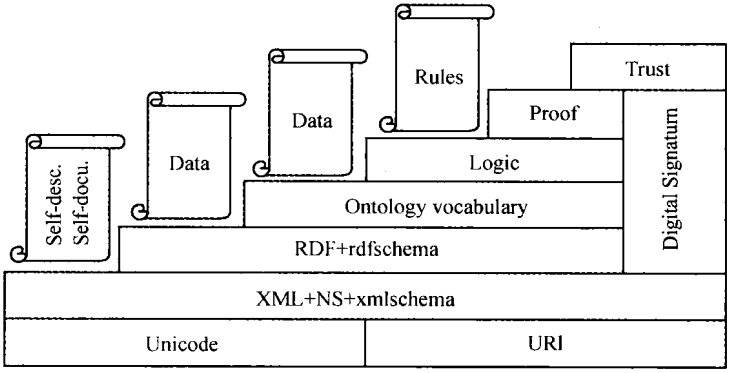

图 1-2　语义网体系结构

语义网在人们与计算机之间实现了语义交流，而本体机制则是实现语义网的关键，是解决语义层次上网络信息共享和交换的基础。

➢　本体

本体原本是哲学中的概念，是对客观存在的系统解释，描述现实的抽象本质[①]。在 20 世纪 90 年代中期，本体被引入知识工程领域，用于描述知识的内涵，表达知识的语义。到目前为止，知识工程领域尚未对本体形成统一定义，研究者在不同实践中应用本体，不断赋予本体以新的内涵。一般认为，本体是共享概念模型的形式化规范说明，包含 4 方面的含义：概念模型、明确性、形式化和共享性[②]。① 概念模型，指它是通过抽象出客观世界中一些现象的相关概念而得到的模型，其表示的含义独立于具体的环境状态。② 明确性，指概念、类型及概念上的相互约束都有明确的定义。③ 形式化，指本体是计算机可读(识别)的。④ 共享性，指本体中体现的是共同认可的统一化的知识，反映的是相关领域中公认的概念集，所针对的是团体而不是个体的共识。

本体是一种元数据，它提供丰富原语描述领域的概念模型，澄清领域

① 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.

② Studer R，Benjamins V R，Fensel D. Knowledge Engineering，Principles and Methods [J]. *Data and Knowledge Engineering*，1998，25(122)：161-197.

知识的结构[①]，具有知识表示的能力，图 1－3 为报纸领域的本体片断；本体可重用，避免了重复的领域知识分析；本体提供了大量受约束的、明确定义的、机器可处理的统一术语和概念，可以构建完整的“术语表”来定义网络中的数据，使知识共享成为可能；本体还能够对知识进行推理和验证。

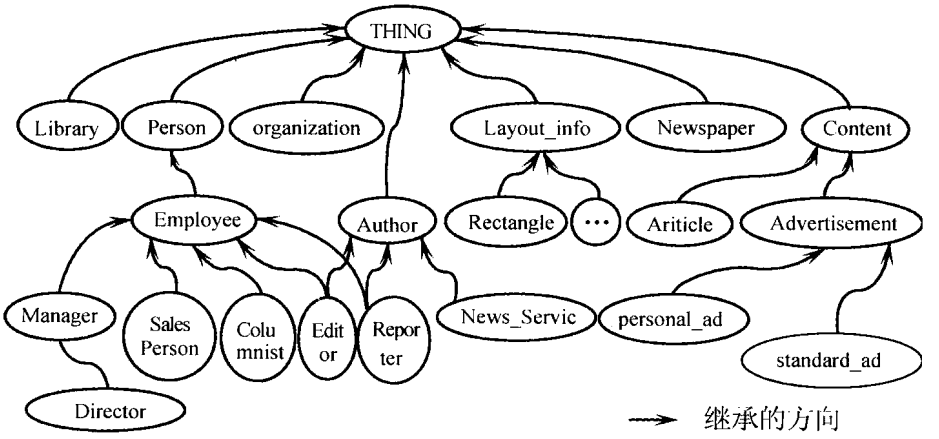

图 1－3 “报纸”本体片断

本体在语义网概念提出之前就已经被引入知识工程中作为知识描述和知识组织的新手段，而语义网的提出，使得作为其实现核心的本体技术得到了全面充分的发展。知识工程各研究领域都对本体技术进行了实验性研究和应用，以探索本体机制的真正价值，并试图建立各种领域本体以作为语义网的资源基础。

随着语义网概念的提出，新的知识组织方式——本体机制得到了前所未有的重视，图 1－4 展示了本体机制在信息组织中所处的地位。作为现代信息领域的新兴技术和有效的知识组织方式，本体还被广泛应用于除语义网外的其他各项研究，如智能信息检索、搜索引擎、电子商务、自然语言处理、软件工程、知识管理、数据挖掘、多代理系统、机器学习、信息分类、地球信息科学和数字图书馆等。

① 汪方胜，侯立文，蒋馥. 领域本体建立的方法研究[J]. 情报科学，2005，23(2)：241－244.

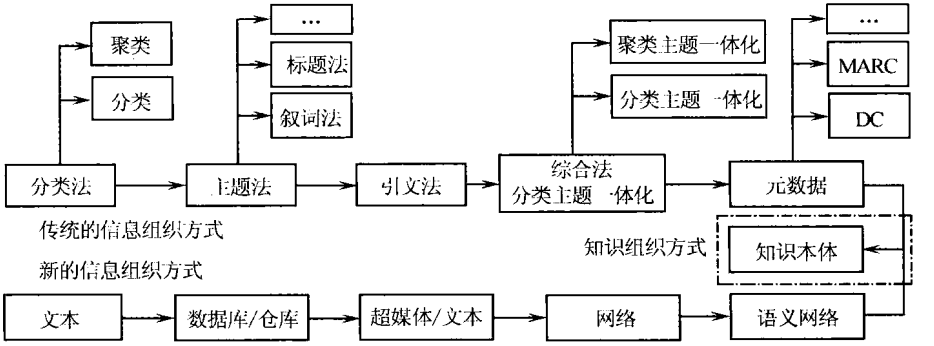

图 1-4　信息组织方式的发展变化

本体机制的广泛研究和应用，使得本体能够明确描述对象间关系的能力获得了普遍的认同，这为进一步充分挖掘和描述 CSSCI 学术资源之间复杂的关系提供了良好契机。CSSCI 自 20 世纪 90 年代末诞生以来，以其规范、权威的检索和分析服务得到了研究者的一致认同。然而，CSSCI 检索服务的简单化和直线型的信息组织使得用户在使用时并不方便，用户在检索目的不太明确的情况下，有时甚至无法获得查找结果。此外，更加精确、能够发掘隐含知识的引文分析目标也对传统的基于简单统计的 CSSCI 分析服务提出了更高要求，研究人员试图通过其他手段如数据挖掘技术等来对 CSSCI 学术资源进行更为深入的剖析，以期获得一些意想不到的结果。在这种情况下，基于本体机制提升 CSSCI 学术资源服务的方案被提了出来。本书试图借助本体对知识的有效组织和对新知识的合理逻辑推理，来解决用户需要更完善的引文服务与 CSSCI 提供的学术资源服务落后之间的矛盾。

(1) CSSCI 学术资源的知识重组需求

CSSCI 学术资源以来源文献和引文文献作为主要研究对象，其他类型对象主要是起辅助描述文献对象的作用，因此现有的数据结构只能反映出其他类型对象和文献对象之间的相互关系，对于其他类型对象相互之间关系，例如主题与作者，主题与机构等等均无法表现；对于同一类对象不同实例之间的关系，例如作者与作者、主题与主题等也无法描述；对于对象之间的关联程度也无法体现。为了使用户了解 CSSCI 学术资源对象间相互关系，提供关联检索服务，充分挖掘和分析学科内在特征，需要使用本体技术来明确描述资源对象之间的关系，实现 CSSCI 学术资源知识重组。

(2) 用户查询表达式引导构建需求

用户在进行信息检索初期往往无法明确自己的检索需求,因此经常获得大量无关信息。借助本体对对象关系的明确描述,可以使用户了解其所要查询对象的具体情况,逐步引导用户明确自己的查询需求。例如用户想查询"关联规则挖掘"主题相关的文献,但在检索初期他并不知道这个术语,只知道它与"数据挖掘"相关,如果用户直接使用"数据挖掘"作为关键字检索,结果得到大量无关信息。这时系统若借助主题本体先返回与"数据挖掘"相对应的主题及与该主题存在继承关系、同义关系、相关关系等的其他主题供用户选择,可逐步引导用户构建需要的查询式"主题=关联规则挖掘";此外,用户根据对与主题相关其他对象如作者、机构、期刊等的了解也可以修正自己的查询式。

(3) 词语匹配向概念匹配过渡实现智能语义检索的需要

传统的基于关键字匹配的信息检索仅考虑到了语法层次的问题,摒弃了检索条件和检索内容的概念特征,信息量损失非常大,导致检索结果质量不能令用户满意。语义检索是一种基于知识分析的智能信息检索,能够对检索条件和检索内容进行了语义层面的处理(语义标引、语义扩展),在自然语言理解的基础上借助数学模型,在关联知识作用下完成概念匹配检索,从而达到更高的查准率和查全率。

本体具有良好的概念层次结构并支持逻辑推理。借助领域本体,可对检索条件和检索内容进行语义分析和语义理解,消除歧义,可采用相应的检索策略和抽取算法实现检索条件和检索内容的概念匹配。在整个检索过程中,以描述领域知识的本体作为基础,可实现智能语义检索,同时在适当情况下还可以进行相关性检索和查询扩展,避免"主题孤岛"问题产生。

(4) 基于本体的引文分析探索

现有的引文分析总是借助对资源某一方面特征的描述,来分析对象之间的关联,例如基于同被引聚类或基于引用或基于主题共现等方式讨论学科、期刊等之间的关联,这种分析存在一定的片面性。本体以面向对象的方式组织领域知识,能够实现对对象的综合描述。借助本体对对象的全面、清晰的描述来进行 CSSCI 学术资源的关联、热点和趋势分析,可以挖掘出更多、更准确的具有参考价值的信息,为用户提供更好的引文分析服务以支持用户决策;能够更加充分地发现学科的内在特征,从而找到促进学科发展的立足点。