C.B.J.J
北京市社会科学理论著作
出版基金资助

# The Knowledge Organization System in Digital Libraries

# 数字图书馆的知识组织系统
## ——从理论到实践

王 军 ◎著

北京大学出版社
PEKING UNIVERSITY PRESS

北京市社会科学理论著作出版基金资助

# 数字图书馆的知识组织系统

## ——从理论到实践

王 军 著

北京大学出版社
PEKING UNIVERSITY PRESS

**图书在版编目(CIP)数据**

数字图书馆的知识组织系统：从理论到实践/王军著. -- 北京：北京大学出版社，2009.1

ISBN 978-7-301-14903-4

Ⅰ. 数… Ⅱ. 王… Ⅲ. 数字图书馆－研究 Ⅳ. G250.76

中国版本图书馆 CIP 数据核字(2009)第 006524 号

**书　　名：数字图书馆的知识组织系统——从理论到实践**

著作责任者：王　军　著

责 任 编 辑：王　华

标 准 书 号：ISBN 978-7-301-14903-4/G・2577

出 版 发 行：北京大学出版社

地　　址：北京市海淀区成府路 205 号　100871

网　　址：http://www.pup.cn

电 子 信 箱：zpup@pup.pku.edu.cn

电　　话：邮购部 62752015　发行部 62750672　编辑部 62752038　出版部 62754962

印　刷　者：涿州市星河印刷有限公司

经　销　者：新华书店

730 毫米×980 毫米　16 开本　13.75 印张　218 千字

2009 年 1 月第 1 版　2009 年 1 月第 1 次印刷

定　　价：30.00 元

## 内容简介

构造机器可以理解的、可自动应用的知识组织系统是数字图书馆的核心研究课题。本书深入地研究了如何利用传统的知识组织资源来实现这一目标。全书由三篇构成：上篇介绍了网络知识组织系统的发展，重点是如何基于传统的知识组织资源（包括分类法、主题词表和元数据）来构造数字图书馆的知识组织系统，以支持概念检索和知识管理等服务；中篇探讨了词表的自动丰富机制；下篇以国际上使用最广泛的杜威十进制分类法为例，深入研究了如何改造图书分类法来实现自动分类。本书在深入浅出的理论分析基础上，给出了许多原创性的解决方案。在真实数据集上所进行的大规模实验和所开发的原型系统充分验证了这些方法的实用性和推广价值。

本书是国内在数字图书馆知识组织系统方面的第一本专著。对从事数字图书馆、知识组织、自动分类、网络信息资源组织等研究和应用开发的科技人员有很高的参考价值，也可作为高等院校信息管理与信息系统、图书馆学与情报学、计算机科学与技术等专业的研究生或高年级本科生的教学参考书和技术资料。

# 序

## 传统知识组织方式同现代信息技术相结合的成功尝试

互联网改变了我们生活、学习、工作和娱乐的方式，已经成为商务、科研和社会交往不可或缺的基础设施，使人类享受到信息时代的文明，它对今天社会信息化的贡献无论再高的评价也不过分。但是人们今天同样无法回避这样的现实：人类产生、收集信息的能力超过了人类组织、管理和有效利用信息的能力，人们急切需要的信息通常被实际不需要的垃圾信息淹没掉。另外，同样不容被忽视的现实：当前互联网技术日新月异，又在进一步改变我们的日常生活，各种业务、商业、市场营销、财务、出版、教育、研究等活动及其工作程序正在得到重新定义，生产、储存、处理、检索和利用信息的方式正在发生新的革命。尽管以互联网为基础的信息系统仍在照常工作，但是它们常常不能提供人们真正需要的东西，却把人们不需要的东西硬塞给人们。网上的内容和内容关系被一成不变的过程紧紧捆绑在一起，现有网络的基础设施和模式给各行各业强行造成了信息和技术断裂。一些前沿的行业正在从战略上面对这些断裂，出现了获取、吸收和共享知识的新模式、创新和经营战略，当前比以往任何时候更需要发展和充分利用“网络智能(Web Intelligence)”的新的高效率的技术和方法，以帮助用户避免不相干的网页搜索结果(网页，链接等)、欺诈的电子商业交易、非个性化网络信息以及错误的网络决策等。

著名的 Springer 科技出版公司计划出版的《计算智能研究》丛书将介绍各个领域计算智能新的发展和进步，意图涵盖工程、计算机科学，物理学和生命科学等领域以及它们背后的支撑方法所涉及的计算智能理论、应用和设计方法。该丛书将包含横贯各个领域在计算智能方面的专著，涉及的领域包括神经网络，联结系统、遗传算法、进化计算、人工智能、元胞自动机、自组织系统、软计算、模糊系统以及混合智能系统等。其中，丛书之一的《网络智能》一书，将探讨新型智能网络理论和下一代利用网络的系统、服务和环境，主题将涵盖：网络信息检索、知识网络和管理、信息管理和数据表现、网络挖掘、网络耕作、网络代理和基于代理的系统、网络安全、信息过滤和访问控制模型、概念本体、语义网、社会智能设计、以网络为基础的支持系统、人网互动、E-应用和系统、万维网技术和协议等。这些课题有助于我们了解网络智能各个方面的技术前沿和最新进展。综上所述，信息技术(Information Technology，IT)所有领域都在不同方向采用不同路径发生变革，没有人能预料我

们图书馆信息服务领域将会出现什么，未来会是个什么模样。

“知识组织”概念是 1929 年英国著名分类法专家布利斯(H. E. Bliss)最早提出的。知识组织是传统主题分析、内容分析的继承和发展，是图书馆学、情报学分类法、标题表、叙词表等标引工具的延伸。知识组织是知识管理的前提，是构筑知识服务大厦不可或缺的基础设施。对知识的编码和有序化决定了知识是否可获得，对知识的有效组织，决定了知识的传播、储存、应用、共享是否可行。互联网问世后，雅虎(Yahoo)、谷歌(Google)、百度等搜索引擎在信息采集、处理和更新的速度、海量检索、即查即得等方面的表现，使传统检索方法和工具在利用网络信息资源面前黯然失色，吸引了传统图书馆信息领域的广大用户。但是搜索引擎也很快暴露出其无用信息泛滥、有用信息贫乏的致命弱点，它们在尝试了无数的工具和方法，才又发现知识组织系统的魅力，并掀起对网络环境下知识组织系统的研究、开发、应用的高潮。知识组织在发生质的飞跃，并在不断衍生许多边缘学科，上述 Springer 出版社《网络智能》一书拟覆盖的范围可以说都是知识组织研究这一主轴下正在出现的各个边缘学科领域。传统知识组织和网络时代的知识组织有共性的东西，都是对人类知识结构进行表达和有组织的阐述的语义工具，但是后者并不是前者的简单的重复或改良，而是螺旋式上升的质的飞跃。

王军的《数字图书馆的知识组织系统——从理论到实践》一书触及当今数字图书馆的核心研究课题，研究了如何改造传统的知识组织工具和方法，使得它们可以应用于数字图书馆和网络环境下，实现数字信息资源的自动组织。

本书研究了基于传统的知识组织资源、开发数字图书馆知识组织系统(Digital Library Knowledge Organization System, DLKOS)需要解决的三大问题：DLKOS 的自动构造，DLKOS 的自动丰富机制，传统知识组织工具的改造。上篇提出了一个通过集成分类法、主题词表和元数据来构造 DLKOS 的方法；中篇介绍了在《中国分类主题词表》和《ADL 地理特征词表》上所做的词表自动丰富研究；下篇研究了如何改造《杜威十进制分类法》来实现书目数据的自动分类。这些工作都是通过分析和挖掘书目数据完成的。书目数据，是知识组织过程所生产的智力产品，其中包含了丰富的领域知识和标引知识。所谓领域知识，是指书目数据中包含的对原始文献内容主题的描述，这是学科领域新词汇、新概念的来源；所谓标引知识，是指书目数据中的分类和标引信息，它们本质上是描述领域知识的自由词汇(如文献标题)到控制词汇(类目、主题词等)之间的映射，其间凝聚了图书馆员的智力劳动。依据从书目数据中挖掘出来的领域知识和标引知识，可以实现传统知识组织工具的改造，DLKOS 的自动构造和丰富。本书的实验都是在来自于现实应用的、大规模的数据集上进行的，包括：中国分类主题词表、杜威十进制分类法(Dewey Decimal Classification, DDC)、美国国会图书馆主题词表(Library of Congress Subject

Headings，LCSH)、北京大学图书馆中文书目数据、美国国会图书馆英文书目数据。实验数据具有常用性、普遍性，实验方法通过多层次的对比和验证以得到最优解决方案，所开发的原型系统在网络上提供公开访问和测试。

本书汇集了作者几年来在这一领域的主要研究成果和劳动结晶，并经过精心组织和编辑，具有研究报告和教科书双重品质。本书既介绍了作者及其团队对于构造机器可理解和可自动应用的知识组织系统的研究成果，而且还系统地论述了网络知识组织系统的发展历史、原理、方法、实现过程和采用的工具等，具有教科书通常具备的全面性、系统性、普遍性的特点。除了深入浅出的理论分析，还介绍了所进行的大量实验和所开发的原型系统，验证了方法的有效性、可行性和实用性。可以说本书是网络知识组织系统领域名副其实从理论到实践的专著。

从整体上看，我国的信息基础结构和科技水平同世界科技先进国家有一定差距，但在局部领域起跑线是同等的。知识组织研究是一个极大的挑战，也是一个绝好的机会。在世界范围内，人人在这一挑战和机会面前是平等的。当前我国经济、科技、文化、教育等正处于盛世时期，我们有理由期待我国的 IT 出版界能够出现 Springer 的《计算智能研究》丛书和《网络智能》这样的专著。我们高兴地看到，王军同志《数字图书馆的知识组织系统——从理论到实践》专著在这样的形势下出现在我们的期待中。本书的研究成果有原创性，开发的原型系统有可借鉴性和实用性，本书提出的改造传统知识组织理论方法使其在数字图书馆或网络环境下应用的技术路线具有很高的学术价值和实用价值，对高等院校信息管理与信息系统、图书馆学与情报学、计算机科学与技术等专业和广大 IT 科技人员都具有普遍的参考价值。知识组织是过去、现在、未来图书馆服务、信息服务、知识服务永恒的主题，知识组织的研究和发展正在向多元化方向发展，参与知识组织的研究、试用、开发、制定标准的机构、组织、行业、人员的规模，都远远超出原来的范围。我们期待有更多的类似王军同志的《数字图书馆的知识组织系统——从理论到实践》这类探索知识组织的专著出现。

曾民族　中国国防科技信息研究中心研究员、学术顾问
中国科技情报学会计算机应用专业委员会顾问

# 前　言

知识组织，又称为信息组织，简单地说就是对信息对象进行整序，以便利信息对象的查找和利用。从简单的文件存放、书籍排序到复杂的文献标引和分类，都是知识组织的例子。在组织的过程中，常常需要参考某种特定的、规范的概念体系和知识结构对信息对象进行规范化的描述和整序，因此这一过程称为知识组织。例如，在图书分类的过程中需要参考公认的图书分类法，它定义了各个类目和类目间的等级和参照关系，可以看作是对所涉及的学科知识结构的描述。知识组织是文献信息环境下人们利用信息的主要方法，它是在图书馆领域形成的管理和利用图书资料的基本手段，是传统图书馆工作的重要基础。但是随着人类信息环境向网络平台的迁移，数字资源成为主流，搜索成为泛在的工具，传统的知识组织理论、方法和工具因囿于传统图书馆的范围、未能在网络信息环境下得以普遍应用而受到质疑，进而动摇了图书馆的存在基础。

本书深入地研究了如何改造传统的知识组织工具和方法，使得它们可以应用于数字图书馆和网络环境下，实现数字信息资源的自动组织。从事这样一项研究，首先要回答的问题是，在搜索引擎一统天下的今天，我们还需要对信息进行组织吗？知识组织这一传统是否已经过时？

观察一下我们日常利用信息的习惯，就可以了解信息组织的方法不可替代。在我们自己的个人电脑中，绝大多数的用户都将个人的电子文档分门别类地存放在文件夹下，而不是随意存放后再装一个桌面搜索工具来解决问题。在网络上，是否有良好的组织结构是影响用户能否顺畅地浏览一个网站的关键因素。考察万维网(Web)信息资源利用方式的发展历史，组织是最早被应用的手段。正是借助对网页系统化的分类组织，雅虎(Yahoo!)成为最早的 Web 信息门户。分类组织，是人类认识世界最基本的思维方法之一，也是人们利用信息的基本手段之一。那么，为什么在十余年的努力之后，Yahoo! 放弃了对雅虎网页目录(Yahoo! Directcry)的大规模系统化维护；为什么组织的方法在网络信息环境下失去了往日的地位，而让搜索技术独领风骚呢？

知识组织的理论与方法在传统的文献信息环境下已发展得相当成熟。正是因为它过于成熟与稳定，未能及时变革以跟上信息环境的急剧发展和变化，未能适应数字化信息资源的特点，以满足网络信息环境下信息资源组织与利用的要求。今天社会大众“即刻满足”的信息消费习惯也使得人们偏好使用搜索引擎，属于“精加

工”的组织方法不再是主流。

但是，当人们需要深入掌握和消化信息时，当信息资源服务于研究和学习的目的时，组织仍然是必不可少的工具，特别是在数字图书馆中。数字图书馆所保存的信息资源是整个社会最宝贵最有价值的信息财富，要充分发挥这些资源的价值，需要对它们进行深度组织。良好的组织，也是向用户提供高质量的、基于内容的、智能的信息服务的前提和基础。搜索技术的进一步发展也需要融合知识组织的方法。今天我们在谷歌(Google)、Ask Jeeve 中用到的词汇辅助功能，Vivisimo 和 MSN Search 提供的结果聚类，为数众多的企业网站为了提高搜索的命中率而在网页中嵌入元数据并提供规范的内容描述，这些都是组织方法补充搜索技术的典型例子。

全面发展网络环境下的知识组织工具和自动组织方法，是数字图书馆的重要研究课题。有两条路线：第一，对传统的知识组织工具和方法进行改造，使之适应网络环境下信息资源组织的需要；第二，发展新的知识组织工具和方法，例如大众分类法(folksonomy)、本体(ontology)。

我们首先来看第一条路线。传统信息环境下的知识组织包括三个部分：① 知识组织工具(例如分类法、词表)；② 知识组织方法(指应用知识组织工具对信息对象进行组织的方法，例如编目规则、分类标引规范等)；③ 知识组织产品(即对信息对象进行组织后生成的二次文献产品，例如书目数据)。相应地，对传统方法进行改造可从三个方面来进行：① 改造传统的知识组织工具；② 实现自动的知识组织过程；③ 利用、挖掘已有的知识组织产品。本书在这三个方面都进行了全面深入的理论分析，提出了创造性的解决方案，在真实数据集上进行了大规模的实验，验证了方法的有效性、可行性和实用性。

对传统的知识组织工具进行改造的目标是使得改造后的网络知识组织系统具有如下三方面的特征：① 是机器可理解的、可交换的；② 具备自动更新、自动丰富的能力，以跟上学科知识更新和发展的步伐；③ 支持对网络信息资源的自动组织，包括自动分类和自动标引等。本书提出的改造传统知识组织工具的方法有三：① 通过集成已有的知识组织资源构造数字图书馆知识组织系统(第三、四章)；② 对词表进行自动丰富(第七、八章)；③ 对分类法进行改造实现自动分类(第十二章)。所有的这些改造方法都是建立在对已有的知识组织产品——书目数据挖掘的基础之上的。如果把知识组织系统看成是抽象的概念体系，那么，书目数据就是概念体系所对应的实例数据。实例数据是新词汇、新概念、新关系、新结构的来源，将这些新的知识挖掘提取出来，根据抽象概念和实例数据间的对应关系对原有的概念体系进行丰富，并根据实例数据的统计分布改造原有的概念体系，这就是本书各项研究的一个基本思路。本书的实验都是在来自于现实应用的、大规模的数据

集上进行的，包括：中国分类主题词表、杜威十进制分类法（DDC）、美国国会图书馆主题词表（LCSH）、北京大学中文书目数据、美国国会图书馆英文书目数据。本书的技术路线是通过多层次的对比试验后确定的最优解决方案，试验结果真实可靠。所有的算法和实验过程都给出了详细的描述，是可重复、可复制的。所开发的原型系统在网络上提供公开访问和测试。

这些研究工作是一个长远研究图景中的第一步。我们的最终目标是实现网络信息资源的自动组织，这要求网络知识组织系统具备从网络信息资源（如网页）中直接学习新知识的能力。网络信息资源相当自由和松散，缺乏一致的结构。这就要求网络知识组织系统具备相当丰富的领域知识才行，简单地说，就是要拥有丰富的词汇和词汇间的关系。传统的知识组织工具中的词汇都是规范的受控词汇，所覆盖的领域知识在广度和深度上都很不够。我们目前所做的，首先是从书目数据中提取新知识来改造和丰富已有的组织工具。书目数据是图书馆员智力劳动的产品，其中的标引信息，反映的是真实数据（图书）到规范概念（主题词、类目）间的映射关系。基于这些明确地、规范地描述出来的映射关系，可以相对较容易地对受控的概念体系进行改造和丰富。完成了这一步，网络知识组织系统便具备了从自由网页中吸收新词汇、新概念和新结构的能力。因为此时，从元数据中提取出来的词汇充当了网页中的新词到知识组织系统中的受控概念间的桥梁。像这样滚雪球般，网络知识组织系统将逐渐丰富、扩充，终将具备处理网络全文资源的能力。在本文写作之际，我们已经在这方面取得了突破性的进展。以 DDC 分类法为分类系统，以国会图书馆的书目数据作为训练数据，我们对 Google Book 的搜索结果进行了自动分类，取得了良好的效果。这验证了本书的技术路线是可扩展的、可推广的，有着现实的应用意义。

在上文提到的第二条路线方面，即发展新的知识组织工具和方法，我们做了两方面的尝试：① 基于传统知识组织资源构建本体 KVision，实现语义检索，并增强搜索引擎；② 构建中文社会书签系统“兜乐”。KVision 语义检索系统基于传统知识组织资源构建本体，基于该本体实现语义查询和知识浏览，并向搜索引擎提供词汇辅助，实现检索结果的聚类。兜乐中文社会书签系统将 Folksonomy 和控制词表结合在一起，在用户端收集用户收藏的书签资源和用户组织这些资源的词汇，在服务器端将书签资源汇集成一个社会书签系统，用收集来的词汇丰富控制词表中的概念和关系，并基于此向用户建议组织本地资源的主题结构、向用户推荐感兴趣的书签资源。KVision 系统在本书的第四章有所介绍；Doulor 尚未包括到本书中来，感兴趣的读者可以访问兜乐网站（www. doulor. com）。这些实践说明，实用的网络组织系统将是两条路线的结合：建立在传统知识组织资源的基础上，结合网络知识组织领域的最新进展，利用搜索、机器学习、挖掘等技术，实现网络环境下的

自动知识组织。

本书的内容是作者长期致力于数字图书馆知识组织这一领域的研究成果的汇总。全书由上、中、下三篇构成，各篇内容既有一定的逻辑联系，反映了作者研究思路的发展脉络；又各自自成体系，可以分别阅读。上篇的重点是基于传统知识组织资源构建数字图书馆知识组织系统(DLKOS)的方法(第二、三、四章)，主要取材于作者的博士论文。其中的重点是第四章，对于理论模型不感兴趣的读者可以略过其他章节。第五章讨论了基于 DLKOS 可以开展的服务，包括概念检索和企业的知识管理。第六章综述了网络知识组织系统这一研究领域的现状和发展趋势。中篇研究了词表的自动丰富机制，介绍了在中国分类主题词表和美国亚历山大数字图书馆的地理特征词表上所进行的卓有成效的实验。第五至八章的内容取材于作者独立撰写以及和学生葛宁、张贝妮、王一丁、林媛合作撰写的若干篇学术论文。下篇以杜威十进制分类法为例研究了如何改造图书分类法以实现自动分类，在杜威十进制分类法和国会图书馆十年的书目数据上所进行的分类实验是迄今为止国内外在这一研究方向上所进行的规模最大的实验。建议时间有限的读者重点阅读第四、六、七、十二章。

感谢北京大学图书馆、美国国会图书馆和联机计算机图书馆中心(Online Computer Library Center，OCLC)为本书的研究无偿提供了实验所需的数据资源。感谢国家自然科学基金(项目编号 70303002)、OCLC/ALISE 联合设立的“情报学图书馆学研究基金”对本书研究工作的资助。OCLC 是世界上最大的提供网络文献信息服务和研究的机构，ALISE(Association for Library and Information Science Education)是全美图书馆学情报学教育联合会。值得一提的是，这是该项基金设立二十余年来首次资助亚洲地区学者的研究工作。感谢北京大学科研部门的领导为本书的研究工作所特批的设备经费。

以此书向在科研道路上授业的诸位老师汇报。在武汉大学陈光祚教授的指导下，作者完成了从事科学研究的学术训练，并培养了科研的兴趣。陈先生数十年如一日孜孜以求地钻研学问，是学生在科学园地里坚持不懈地耕耘的榜样。在北京大学杨冬青教授和唐世渭教授的指导下作者开始探索数字图书馆这一领域，奠定了本文的研究基础。感谢国防科技信息中心的曾民族教授长期以来对作者的鼓励，并为本书作序。作者衷心感谢在 KVision 课题组里先后工作过的学生：葛宁、祖勇、邓鹏、朱兴国、张丽、张贝妮、崔健海、程煜华、税敏、李孟臣、王一丁等。他们参与了本书中的各个实验，承担了大量的编码工作，开发了数个原型系统。研究的思想受益于作者和他们的讨论，课题的成果离不开他们的聪明才智和辛勤努力。回顾作者多年的教学和科研经历，最大收获莫过于曾经和这样一批优秀的学生一起工作，一同前进。最后，感谢北京大学出版社的杨立范主编诚挚的帮助，感谢沈

承凤编辑和王华编辑认真负责的工作，感谢李孟臣同学协助校对了下篇的文稿。

由于时间所限，一些最新的研究成果尚未整合进来，例如基于传统知识组织资源对 Google 的检索结果进行分类、结合 Ontology 和 Folksonomy 设计的社会书签系统“兜乐”。感兴趣的读者请关注 KVision 网站，以了解最新研究进展。由于写作仓促，文中错误在所难免。希望能得到热心读者的反馈，使得本书的研究工作臻于完善。

作者

2008 年 9 月于北大中关园

# 目　录

## 上篇　集成传统知识组织资源，构造数字图书馆的知识组织系统

## 中篇 词表的自动丰富机制研究



## 下篇 图书分类法自动分类研究

# 上篇

## 集成传统知识组织资源，构造数字图书馆的知识组织系统