山西省科技基础条件平台项目资助
“山西省科技文献共享与服务平台管理与利用机制研究”成果

山西科技文献资源整合与数据挖掘技术研究

尚成国　张国瑜　等◎著

技术文献出版社

山西省科技基础条件平台项目资助
“山西省科技文献共享与服务平台管理与利用机制研究”成果

山西科技文献资源整合与数据挖掘技术研究

尚成国　张国瑜 等◎著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS
·北京·

图书在版编目(CIP)数据

山西科技文献资源整合与数据挖掘技术研究/尚成国等著. —北京：科学技术文献出版社，2016.12

ISBN 978-7-5189-2137-9

Ⅰ.①山… Ⅱ.①尚… Ⅲ.①计算机网络—应用—科技文献—文献资源建设—研究—山西②数据收集—研究—山西 Ⅳ.①G253-39②TP274

中国版本图书馆 CIP 数据核字（2016）第 282756 号

山西科技文献资源整合与数据挖掘技术研究

策划编辑：周国臻 责任编辑：李 晴 责任校对：赵 瑗 责任出版：张志平

出 版 者 科学技术文献出版社
地 址 北京市复兴路15号 邮编 100038
编 务 部 （010）58882938，58882087（传真）
发 行 部 （010）58882868，58882874（传真）
邮 购 部 （010）58882873
官方网址 www.stdp.com.cn
发 行 者 科学技术文献出版社发行 全国各地新华书店经销
印 刷 者 北京九州迅驰传媒文化有限公司
版 次 2016 年 12 月第 1 版 2016 年 12 月第 1 次印刷
开 本 880×1230 1/32
字 数 190千
印 张 7.5
书 号 ISBN 978-7-5189-2137-9
定 价 42.00元

序

科技文献资源作为科技发展的基础性资源，其重要性不言而喻。随着信息技术和网络技术的迅速发展，科技文献的生产、存储和传递方式发生了本质性的变化，数字科技资源逐渐成为科技文献资源的主体，对科技文献资源的深度开发和充分利用已成为现代科技进步和社会经济发展的重要推动力和关键要素。在当前激烈的国际竞争中，其重要性日益凸现。20世纪90年代美国政府把数字信息资源生产、传播和利用作为国家信息化建设的关键和重点。1998年欧洲启动了面向数字信息资源的整合开发项目。2002年，加拿大在提出的国家创新体系中，将建立国家数字科技信息网作为重要组成部分。世界各国政府对数字资源建设与利用都给予高度重视，为抢占科技发展的制高点，纷纷投入巨资将数字资源平台建设作为提升本国科技创新能力的重要手段。

近年来，我国政府从增强国家自主创新能力的核心竞争力的高度出发，对科技文献资源建设非常重视。中共中央办公厅、国务院办公厅印发的《2006—2020年国家信息化发展战略》，明确提出“大力发展以数字化、网络化为主要特征的现代信息服务业，促进信息资源的开发利用”，这充分表明了国家对数字资源建设的重视程度。科技文献资源作为信息资源的主体，是国家的战略资源，是国家科技创新的重要支撑和基础保障。为了加强科技文献资源共享平台建设，国务院办公厅转发了由科技部、国家发改委、教育部、财政部联合制定的《2004—2010年国家科技基础条件平台建设纲要》(以下简称《纲要》)。《纲要》明确指出，到2010年要初步建成适应科技创新需求和科技发展需要的科技基

础条件支撑体系，以共享为机制核心的管理制度，与平台建设和发展相适应的专业化人才队伍和研究服务机构，为最终形成布局合理、功能完善、体系健全、共享高效的国家科技基础条件平台奠定基础。《纲要》中提到的科技基础条件重点建设的六个平台，其中科技文献资源共享平台就是主要平台之一。2005年山西省科技文献共享与服务平台就是在这个大背景下开始启动建设的，现已初步建成并对社会开放。“十二五”期间，根据科技部的要求，科技文献共享平台的重心要由建设逐渐转向利用与服务，因此本项目组承担了山西省科技基础条件平台项目“山西省科技文献共享与服务平台管理与利用机制研究”。本项目以国家和山西省科技文献资源共享发展战略为指导思想，从山西省科技文献资源战略层面、技术应用层面和管理与利用机制层面，系统论述了科技文献资源建设与服务的基础理论、山西省科技文献资源战略规划；分析了山西省科技文献共享与服务平台的建设内容，研究了山西省科技文献整合与数据挖掘、云服务与云安全；探讨了山西省科技文献共享与服务平台的管理机制与服务机制，对平台的管理与服务进行绩效评价。本项目最终形成了系列研究著作3部，即《山西科技文献共享与服务平台管理及利用机制研究》《山西科技文献资源云服务与云安全研究》《山西科技文献资源整合与数据挖掘技术研究》。本项目研究的特点主要体现在：

第一，从共享视角，将山西省科技文献共享与服务平台视为一项社会科技创新中的系统工程。山西省科技文献共享与服务平台所涉及的资源规划、资源开发、资源管理、资源利用与服务绩效评价不是一项孤立的工作，而是与山西省科技进步和社会经济发展密不可分的关键环节。本项目研究不仅仅关注科技文献资源共享的各个要素，尤其重视平台整体功能的发挥和各种要素的协调、重视平台管理与服务有效机制运行的研究，最终建立一个布局合理、功能完备、高效开放的山西省科技文献共享与服务平台，更好地为用户服务，提高科技文献的利用率。

第二，从云理论的角度，对山西省科技文献共享与服务平台

的构建、服务提供和平台安全进行了分析，并系统阐述云计算的基本理论与基本技术，将云计算应用于山西省科技文献共享与服务平台，利用云的关键技术，构建了平台的云服务和云安全方案，对有效提升平台管理与利用的效益提供了技术理论基础。

第三，从信息技术的角度对山西省科技文献资源整合与数据挖掘技术进行了深入的研究。资源整合与数据挖掘技术研究不仅是技术方法与手段层面的研究，更是对山西省科技文献共享与服务平台管理与利用机制运行更为深刻的研究。本项目综合多种因素，对山西省科技文献资源管理与服务的业务、资源、技术、保障条件等进行多角度的探讨，通过资源整合、数据挖掘，实现跨库、跨平台一站式检索，用户可以方便快捷地获取所需资源，无论从平台管理理念、技术应用还是从利用机制实现等方面都体现出山西省科技文献共享与服务平台在资源整合与数据挖掘技术的实践的特色。

本系列书在资料收集与撰写过程中得到了山西省科技厅科技基础条件平台项目经费的资助，得到了山西省科学技术情报研究所的全力帮助，得到了山西财经大学科技处的大力支持，在此表示衷心的感谢！本系列书在撰写过程中吸收了国内外大量的研究成果，参考和引用了许多学者的有关著述和论文，在此表示真挚的感谢！本系列书的出版是我校情报学教学科研团队导师、信息管理学院教师、图书馆专业人员研究的成果，在此对治学严谨、辛勤探索的各位作者表示感谢！对科学技术文献出版社的责任编辑周国臻先生在策划编辑过程中付出的艰辛表示感谢！同时，希望广大读者不吝指正，共同推动科技文献共享与科技文献平台利用的深入！

山西财经大学信息研究所所长　武三林
山西财经大学信息管理学院院长　贾　伟
2014年12月

前 言

科技文献资源是国家的重要战略资源，是科技创新的重要条件之一。随着科学技术的发展和信息的不断增长，科技文献资源也越来越丰富，对科技文献资源进行管理和利用机制研究，有利于提高科技文献资源的利用水平，为社会发展和科技进步提供重要保障。

山西省科技文献共享与服务平台是山西省科技基础条件平台的重要组成部分。“山西科技文献资源整合与数据挖掘技术研究”为2011年承担的山西省科技厅科技基础条件平台项目“山西省科技文献共享与服务平台管理与利用机制研究”（2011091001-0101）的子课题之一，主要针对山西省科技文献共享与服务平台在科技文献资源异构检索、科技文献资源整合及数据挖掘技术在科技文献资源中的应用等问题进行研究。

本书对山西省科技文献资源整合与数据挖掘问题的研究意义和作用进行了说明，介绍了科技文献资源整合涉及的基本概念、整合模式和相关技术，以及数据挖掘的基本理论和常用技术。探讨了山西省科技文献资源异构检索的方法，分析了山西省科技文献资源异构检索的需求与功能。结合山西省科技文献共享与服务平台，详细阐述了山西省科技文献资源整合方案，包括资源整合架构、主要技术、平台设计等，对数据挖掘技术在山西省科技文献共享与服务平台中的应用进行了进一步的分析。

本书包含7章内容，具体包括山西省科技文献资源整合与数据挖掘技术研究的意义与作用、科技文献资源整合的基本理论、数据挖掘的基础理论、数据挖掘技术、山西省科技文献资源异构检索需求与功能、山西省科技文献资源整合、数据挖掘在山西省科技文献资源共享与服务平台中的应用。

本书第1章由尚成国撰写，第2章至第7章由张国瑜撰写，张国瑜和尚成国负责全书统稿，研究生王宇、赵丽佳、李鑫参与了本书的资料收集工作。

本书获得了项目负责人武三林研究馆员、山西财经大学信息管理学院贾伟教授的指导和帮助，在此表示衷心的感谢！在本书的编写过程中参考了大量的国内外出版物和网上资源，引用了许多专家、学者的有关著述，在此致以诚挚的谢意!对科学技术文献出版社的周国臻、李晴编辑在编辑过程中付出的辛勤劳动，表示由衷的敬意和感谢！

由于著者水平及可获得的资料有限，书中难免出现问题，欢迎专家、读者批评指正。

著　者

2016年10月

目 录
Contents

第 1 章

山西省科技文献资源整合与数据挖掘技术研究的意义与作用

山西省拥有丰富的科技文献资源，2005年起，山西省启动了山西省科技基础条件平台建设工作，山西省科技文献共享与服务平台建设项目是山西省科技基础条件平台建设的项目之一。在此基础上，对山西省科技文献资源进行整合及对数据挖掘技术在科技文献资源中的应用进行进一步的研究，有利于更好地发挥科技文献资源价值，对于促进社会科技、经济发展具有重要意义。

本章阐述了国内科技文献共享平台的建设状况，对山西省科技文献资源整合及科技文献中数据挖掘技术的研究意义与作用进行了分析。

1.1 研究背景

随着全球化的发展和知识经济的到来，为科技创新提供支撑的科技基础条件平台成了国家赢得国际竞争的重要基础保障，科技文献资源共享平台是科技基础条件平台的重要组成部分，它为海量的科技文献资源提供了一个加工、整合的渠道，给科研单位、政府、企业及个人用户提供了一个统一、全面的科技文献资源获取窗口。进行科技文献资源共享平台建设，关键是要进行科技文献资源整合，提高文献资源利用水平，将分散在各单位、异构的文献资源进行有机融合，并在此基础上对科技文献资源进行高效利用，这是推动科技文献资源共享平台建设的主要动力。

科技文献资源是国家的重要战略资源，是科技创新的重要条

件之一。在我国科学技术部（下文简称科技部）出台的《2004—2010年国家科技基础条件平台建设纲要》中，将科技文献共享平台建设列为科技基础条件平台的重要内容，同时指出要建立科技文献共享平台，扩大科技文献资源的收集和服务，为科技文献资源整合共享提出了指导方针。

1.1.1 国内科技文献共享平台的建设状况

资源整合是实现科技文献共享的基础。中国高等教育文献保障系统（China Academic Library & Information System，CALIS）和国家科技图书文献中心（National Science and Technology Library，NSTL）是我国文献资源共建共享的典型范例，在促进我国文献资源整合方面发挥了积极的作用。

中国高等教育文献保障系统，是经国务院批准的我国高等教育“211工程”“九五”“十五”总体规划中的3个公共服务体系之一。该平台在高校图书馆范畴内进行了资源的整合。其宗旨是在教育部的领导下，把国家投资、现代图书馆理念、先进的技术手段、高校丰富的文献资源和人力资源整合起来，建立以中国高等教育数字图书馆为核心的教育文献联合保障体系，实现信息资源的共建、共知、共享，以发挥最大的社会效益和经济效益，为中国的高等教育服务。CALIS管理中心设在北京大学，管理中心下设立文理、工程、农学、医学4个全国文献信息服务中心，华东北、华东南、华中、华南、西北、西南、东北7个地区文献信息服务中心和1个东北地区国防文献信息服务中心，建成了由众多图书馆参与的高校文献“全国中心—地区中心—高校图书馆”三级保障体系和包括文献获取环境、参考咨询环境、教学辅助环境、科研环境、培训环境和个性化服务环境在内的六大数字服务环境的中国高等教育数字图书馆。

国家科技图书文献中心是科技部于2000年6月12日，根据国务院批示组建的虚拟的科技文献信息服务机构，成员包括中国科学院文献情报中心、国家工程技术图书馆（中国科学技术信息研究所、机械工业信息研究院、冶金工业信息标准研究院、中国化工信息中

心）、中国农业科学院图书馆、中国医学科学院图书馆。网上共建单位包括中国标准化研究院和中国计量科学研究院。NSTL实质上是对理、工、农、医科研院所图书馆的资源整合，其宗旨目标是根据国家科技发展需要，按照“统一采购、规范加工、联合上网、资源共享”的原则，采集、收藏和开发理、工、农、医各学科领域的科技文献资源，面向全国开展科技文献服务，最终发展目标是建设成为国内权威的科技文献信息资源收藏和服务中心、现代信息技术应用的示范区、同世界各国著名科技图书馆交流的窗口。NTSL主要任务是通过统筹协调，较完整地收藏国内外科技文献资源制定数据加工标准、规范，建立科技文献数据库，利用现代网络技术，提供多层次服务，推进科技文献资源的共建共享，组织科技文献资源的深度开发和数字化应用，开展国内外合作与交流。

在国家的政策性扶持及引导下，全国各省、市都在建设各自的科技文献共享平台，目的是将可支配的分散资源进行整合，为科技创新做好文献支撑服务。大部分科技文献共享平台都是由科技信息资源丰富的省情报研究机构以及区域内各高校图书馆共建的。已建成的科技文献共享平台中，大多以整合自建数据库、引进数据库、特色数据库为主来丰富和充实平台的信息资源。由于数据库种类繁多，大部分科技文献共享平台都会提供跨库检索功能，方便用户检索信息资源。另外，各省、市科技文献共享平台开展了多样化的服务项目，主要集中在委托检索、原文代查、订阅推送、定题跟踪、科技查新、科技评估等几大项。随着数据挖掘技术的发展，许多平台也纷纷增加了科技文献共享平台的新功能，为用户提供推送服务、自助服务等个性化的服务内容，不断完善平台功能，积极提升用户体验。

1.1.2　山西省科技文献共享与服务平台的建设状况

2004年7月，国务院办公厅转发了由科技部等四部委联合制定的《2004—2010年国家科技基础条件平台建设纲要》，至此，旨在促进我国科技进步、科技创新的国家科技基础条件平台的工

作开始启动。2005年8月，《“十一五”国家科技基础条件平台建设实施意见》发布，明确指出我国“十一五”期间国家科技文献平台的重点建设内容。山西省从2005年起设立专项计划，启动了山西省科技基础条件平台建设工作。

目前，山西省已建有山西省科技文献共享与服务平台，是山西省科技基础条件平台的重要组成部分。该平台于2005启动，是以山西省科学技术情报所作为项目牵头单位来组织和实施的，其他协作单位有山西大学、山西农业大学、太原理工大学、山西医科大学四家省内知名高校。2006年平台项目滚动进行，为满足资源建设的需求，新吸纳山西大同大学、山西财经大学作为试点加入平台建设的行列。2008年，山西省科技文献共享与服务平台建设的成员单位基本涵盖了山西省科技情报机构、高校、科研院所、公共图书馆等重点科技文献持有单位。

该平台的建设是在调查研究的基础上，结合山西省经济建设和科技发展的趋势，利用现代信息网络技术，经过对科技文献资源进行整体规划和设计，采取滚动建设、不断完善的方式来实施，对科技报告、科技成果、会议论文、专利、科技图书、科技期刊、技术标准、工艺等文献资源，进行整合、重组与优化，具备了服务中心、用户指南、统一检索、网站导航等栏目。随着技术的进步和山西省科技文献资源建设工作的发展，山西省科技文献整合平台需要增加新功能、注入新活力，才能满足用户对科技文献不断变化的信息需求。

1.2 研究意义与作用

1.2.1 科技文献资源整合研究的意义与作用

（1）提高科技文献利用率

随着科学技术的发展和信息的不断增长，科技文献资源也越来越丰富。由于科技文献产生的渠道多样，日积月累，大量的科技文献资源处于一种分散、无序的状态，给人们的获取、利用造

成了困难。通过对科技文献资源整合，将分散在图书馆、信息机构、网络上的资源进行科学组织，在对科技文献资源的鉴别、筛选、采集、加工、处理等一系列工作中，将科技文献资源进行有组织地描述、揭示，不但把零散的资源组合为有序的资源，还将各种不同类型的科技文献资源（如多媒体、文本、纸质、胶片等）进行整合，最后通过检索技术把相关内容完整地呈献给用户，从而充分发挥科技文献自身价值。

（2）提高科技文献服务质量

通过数字化的科技文献整合，不但使信息提供机构拥有了更加丰富、有序的科技文献资源，还为用户提供高质量文献服务提供了可能。海量的数字化资源，使得信息提供机构可以充分利用计算机技术、网络技术、人工智能等技术，为用户提供不同层次、不同内容的用户服务，甚至进行服务创新，如信息检索服务、数字参考咨询服务、信息推送服务、个人定制服务、个人图书馆、数据统计分析等。

（3）降低信息提供机构采购成本，提高管理和协作水平

图书馆和图书情报服务机构拥有大量的科技文献资源，尤其是图书情报服务机构，除了拥有常规的图书、期刊资源，还有特色科技文献资源和行业内的专业数据库。这些科技文献资源是科研院所、高等院校及企业的科技工作者，在科研、教学、生产工作中不可或缺的。但这些科技文献资源归属于不同机构、部门，只为本单位用户使用，如果本单位没有，只能通过购买实现，导致资源重复配置。通过科技文献资源整合，可以把区域范围内各个图书馆及图书情报服务机构的资源进行集中，实现共享，降低各单位的资源采购费用。参与整合的多个信息提供机构之间还可以建立业务合作，联合为用户提供各自擅长领域的服务。

1.2.2　科技文献数据挖掘技术研究的意义与作用

平台要想提高资源利用率及服务质量，最重要的是根据用户的需求提供准确、全面、个性化的服务，而数据资源挖掘技术为

解决用户的科技文献需求问题提供了解决方法，它使人们对科技文献的应用从单纯的收集、整理、组织、存储、传播、使用，发展到对科技文献的重建、集成及知识创新。在使用数据挖掘技术对科技文献资源进行分析处理之后，将会从海量的资源数据和业务数据中发现隐含的规律、特征，并将所得的规律、特征运用到用户服务与资源建设的实际工作中，可以使科技文献资源的用户服务工作更有针对性，科技文献资源建设决策也更加科学化、合理化。因此，平台科技文献数据挖掘无论对于满足用户需求还是科技文献资源建设，都具有十分重要的意义。

（1）为用户提供更加优质的服务

在现代信息社会中，科技文献信息越来越多，增长速度越来越快，人们不可能将所有的科技文献全都阅读完毕，同时在庞大的平台科技文献资源中，对于特定的某个用户，并不是所有的科技文献都是有价值的，即使是从中选出更符合用户需求的资源也是很困难的。如何在海量科技文献中选择有价值的文献，已经成为所有读者需要面临的问题。

另外，随着用户需求的多样性和个性化，简单的科技文献搜索服务已不能满足用户的全部需求，用户还需要山西省科技文献共享与服务平台能够主动提供用户感兴趣的文献，能回答用户复杂的、专业的咨询。

在数据挖掘技术的支持下，山西省科技文献共享与服务平台能够发现用户需求特征，预测用户行为，提高用户选择信息的效率，为用户提供个性化、主动式的服务。

①数据挖掘技术可实现用户服务个性化。例如，山西省科技文献共享与服务平台的用户来自于各行各业，包括科研工作者、教师、学生、企业管理人员、政府决策人员、农民、新闻工作者等。不同用户有着不同的特点和需求，如用户的受教育程度、专业、信息素养、习惯爱好、对科技文献需求的层次都有差异，面对这些用户，平台提供的科技文献服务也不应该是一成不变的，必须根据用户的特征和需求，为每一名用户提供独特的个性化服务。

②数据挖掘技术可实现用户服务主动化。在互联网时代，用户获取科技文献的行为方式发生了很大的变化，用户更加倾向于随时随地上网获取信息，平台科技文献资源服务也不例外，只要有网络，用户不需要再到实体图书馆获取资料。用户能否获得好的科技文献资源，取决于用户对所涉及问题领域的了解程度、对问题表述的准确性等因素。如果科技文献资源的用户服务变等待用户提问的被动做法为主动推送信息的服务方式，使用数据挖掘技术从平台科技文献数字资源的业务数据（如用户以往的浏览内容、下载记录等历史信息）中，发现用户使用科技文献资源的规律、关心的主题，结合学科知识的有关内容，可适时主动地向用户推送相关的数字资源，协助用户获取更准确的文献资源，或提供相关信息，方便用户更准确地描述问题。

（2）有助于科技文献资源建设

数据挖掘技术在平台的科技文献资源建设中的作用主要体现在两个方面：一方面体现在资源建设的决策中；另一方面体现在资源优化配置中。这两个方面都是平台科技文献资源建设的重要工作，关系到平台的未来发展。

①辅助平台决策。数据挖掘技术能通过对山西省科技文献共享与服务平台业务数据的处理中获得规律，使科技文献资源建设的各项决策是在强有力的数据支持下产生的，这样的决策会更具有科学性，更能体现科技文献资源服务的个性化特征，更符合平台用户的实际需求。

②优化资源配置。通过对数据仓库中数据的分析，可以得到各类科技文献资源的利用状况。根据科技文献资源的实际使用情况进行文献资源采购，使科技文献结构更加合理，更能符合读者的实际需求，节约了大量的资金，使资源得到了优化配置。如通过对数据的分析，可以了解到文献的利用状况，对于利用率低但有价值的资源通过推荐服务介绍给有需求的用户，对于低价值的资源可以淘汰，从而提高了平台科技文献资源的利用效率。