中国文化典籍计算机整理与开发技术研究系列
丛书主编◇侯汉清

GUJI JISUANJI
XINXI MENHU ZIDONG
GOUJIAN YU YINGYONG
YI NONGSHI XUEKE WEI LI

古籍计算机信息门户自动构建与应用

——以农史学科为例

刘　竟◎著

安徽师范大学出版社

中国文化典籍计算机整理与开
丛书主编◇侯汉清

GUJI JISUANJI
XINXI MENHU ZIDONG
GOUJIAN YU YINGYONG
YI NONGSHI XUEKE WEI LI

古籍计算机信息门户自动构建与应用

——以农史学科为例

刘　竟◎著

安徽师范大学出版社

责任编辑：汪碧颖　　责任校对：潘　安
装帧设计：丁奕奕　　责任印制：郭行洲

图书在版编目（CIP）数据

古籍计算机信息门户自动构建与应用：以农史学科为例/刘竟著．—芜湖：安徽师范大学出版社，2013.11
（中国文化典籍计算机整理与开发技术研究系列/侯汉清主编）
ISBN 978-7-5676-0998-3
Ⅰ.①古… Ⅱ.①刘… Ⅲ.①农业史—古籍—文献信息—研究 Ⅳ.①G257.39

中国版本图书馆 CIP 数据核字（2013）第 238916 号

古籍计算机信息门户自动构建与应用
——以农史学科为例
刘竟 著

出版发行：安徽师范大学出版社
芜湖市九华南路 189 号安徽师范大学花津校区　　邮政编码：241002
网　　址：http://www.ahnupress.com/
发 行 部：0553-3883578　5910327　5910310（传真）　E-mail：asdcbsfxb@126.com
经　　销：全国新华书店
印　　刷：安徽芜湖新华印务有限责任公司
版　　次：2013 年 11 月第 1 版
印　　次：2013 年 11 月第 1 次印刷
规　　格：700×1000　1/16
印　　张：11.5
字　　数：159 千
书　　号：ISBN 978-7-5676-0998-3
定　　价：26.00 元

出版说明

一

中国文化典籍是中华民族在数千年历史发展过程中创造的重要文明成果，蕴含着中华民族特有的精神价值、思维方式和想象力、创造力，是中华文明绵延数千年的历史见证，也是人类文明的瑰宝。对古籍的整理、保护与开发，是中华儿女应尽的义务和职责。

我国古籍资源数字化工作起步于20世纪80年代初期，经过几十年的发展，已取得令人瞩目的成就。第一批《国家珍贵古籍名录》和全国古籍重点保护单位的申报工作早已完成，制定古籍数字化标准列入议程，古籍整理与保护工作进入一个新的历史阶段。

古籍资源数字化最初主要是制作书目数据库，后来发展到古籍全文数据库，直至如今的网络检索系统。信息技术的发展和数字化成果的不断涌现，对古籍数字化提出了更高的要求。专家认为，数字化的古籍资源除了实现文本字符的数字化、具有基于超链接的浏览阅读环境和强大的检索功能外，还需具有“研究支持功能”。所谓“研究支持功能”，是指能够提供有关古籍内容本身科学、准确的统计与计量信息，提供与古籍内容相关的参考数据、辅助工具。这些信息、数据或工具都是古籍内容的增值或补充。北京大学计算语言研究所和古文献研究所合作开发了“古诗研究计算机支持系

统”，并取得了阶段性成果。

时值古籍数字化研究日新月异、如火如荼之际，安徽师范大学出版社于2011年精心策划、2012年成功申报、2013年落实出版国家出版基金项目“中国文化典籍计算机整理与开发技术研究”（编号：2013G2－011），在数字化古籍诸项功能特别是“研究支持功能”上给予探索。

改革开放30多年来，国泰民安、政通人和，中国传统文化日益受到政府重视，有关科研机构加大了对古籍整理研究的力度。安徽师范大学出版社能够有机会申请到国家出版基金项目的资助，本项目丛书能顺利进行，实在与国家关注出版事业、关注中国传统文化、关注文化典籍计算机整理工作密切相关。

二

“中国文化典籍计算机整理与开发技术研究”项目主要内容如下：

第一，探索与试验古籍知识库、模式库，将之改造为规则库。

本项目利用命名实体识别、词汇同义词关系的识别、文本主题概念的提取等技术，从各类古籍数据库抽取人名、地名、文献名、职官名、物品名、年号等；并将人名表、地名表、书名表、年代年号表等，与引书模式、异名别称模式、断句模式、分类模式等模式库整合成一个古籍整理与开发专用的知识库，以方便中文古籍整理与开发。

本项目构建的各类知识库，具体有：古代官名、人名和地名表；避讳字、异体字和繁简字对照表；常用古籍名称库；专业术语词典，按专业分为历史、天文、农业、医学、宗教等多个专业词典；主题术语词典，按主题分为动物、植物、矿物等若干主题词

典；古代关联词语表，用语义相似度计算和基于词典释义的同义词识别算法，开发古代关联词语表；禁用词典。

本项目构建的各类模式库有：异名别称模式库，包括别称词、避忌特称、地域特称、文献特称等；断句标点模式库，包括句法特征词法、同义语标志词法、反义复合词、引书标志、时序、数量词、重叠字词、动名结构及比较句法等多种模式识别库；古籍分词模式库。大多数古籍文本无标点，分词的长度及方法需要单独构建。

这些知识库与模式库，采用拿来主义，并经过计算机检验与筛选，最终形成适用于计算机处理古籍的规则，合成为一个综合的规则库，从而为计算机处理古籍提供有力的规则支撑。

第二，重点探索与试验下列古籍智能整理与开发的关键技术。

自动校勘技术：采用对校法，借鉴中文文本自动校对和模式匹配技术，通过比对程序校勘古籍。

自动断句标点：对现有部分标点本古籍进行数理统计，归纳、总结其断句和标点模式。同时结合语言学方法，进一步优化断句和标点模式，从而实现计算机辅助断句与标点。

自动分词和标引：利用汉语现代文本的分词理论和方法，探索古籍文本的自动分词技术，并利用统计学方法（N-gram 等），从古籍数据库中筛选出有一定表达意义的实词词汇。同时利用异名别称模式，创建并完善古籍用词同义词典。在此基础上，引入文本数据挖掘、主题提取和自动分类技术，探索基于知识库的古籍文本的自动标引与分类。

自动编纂：让计算机模拟人脑从大量古籍文本中判断、选择出与编纂主题相关的资料，实现古籍专题资料的自动编纂工作。

自动注释：收集已有古籍专业词汇及其注解，构建古籍语词注解知识库。

第三，在上述基础上，将它们整合为计算机整理与开发古籍的

“一条龙”服务，即构建出古籍整理与开发的专家系统或智能处理系统。

将以上各种词汇、知识、模式整合起来，构建成一个内容丰富、功能多样的古籍规则库，再与自动校勘、自动断句标点、自动分词标引、自动编纂、自动注释等各项技术结合，从而实现文化典籍整理与开发的“一条龙”服务，提出并设计一种集成各种古籍整理与开发智能技术的原型系统。该系统集知识与模式于一身，集规则与技术于一体，具有合成性，既适用于古籍数据库的建设，又适用于古籍数据库的开发使用。

第四，在上述基础上，本研究进行四项个案研究，在实践中探索上述集成的古籍整理与开发智能技术原型系统的可行性与应用性。

农业历史文献数字化：构建农史文献资源库，对农史文献进行自动标引和自动分类，提供农史文献的浏览与检索服务。

建立农史文献门户：构建农史门户网页智能搜索引擎和农史网页自动标引与自动分类实验系统，构建农史门户实验网站。

探索民国农业文献自动索引：在民国农业文献数字化整理中的具体应用，研究索引自动编纂、电子图书编纂、电子索引编纂、数据库建设和主题网关构建等技术方法。

地方志中农业资料的挖掘：从《方志物产·广东》中选取比较实用的全文数据库、物产索引、引书索引、物产分析和引书分析等几个方面进行研究。

总之，本项目充分利用目前在现代汉语文本已经取得成功的中文信息处理技术成果，并根据此成果中的模式识别技术、聚类技术、信息自动提取、信息检索及其他自然语言处理技术等，对照现已建成的大量数字化文化典籍数据库，归纳并修订各类知识库与模式库，研究古籍的自动校勘、自动断句标点、自动分词标引、自动

编纂、自动注释等技术，合成古籍整理与开发的专家系统或智能处理系统，从而为大规模建设新的更多古籍数据库作准备。

三

本项目成果的推广和运用，不但对于探索数字时代古籍文本自然语言处理的理论和方法具有一定意义，而且对推动古籍整理和研究的自动化和智能化、促进我国文化典籍资源的建设和开发以及弘扬传统文化等方面，均具有重大的现实意义和很高的应用价值，可以为继承与发扬中华古籍文化、为建设中国特色社会主义文化服务。

本项目丛书主编由南京农业大学信息科技学院博士生导师侯汉清教授担任。侯先生是中国古籍整理专业第一个硕士研究生，早年在北京大学任教，现执教于南京农业大学，系中国古籍整理专家、中国索引学会副理事长。中图分类法就是侯先生主创起来的。2008年，侯先生主持国家社会科学基金重点项目“文化典籍整理与开发智能技术研究”（编号：08ATQ002），本套丛书即此项目的纸质成果。

本丛书分为六册，各册的内容及其撰写者简要介绍如下：

《古籍计算机自动断句标点与自动分词标引研究》，侧重于自动断句标点、自动分词标引研究，兼顾古籍计算机整理与开发系统的构建与集成。作者黄建年，博士，研究馆员，现就职于南京财经大学。

《古籍计算机自动校勘、自动编纂与自动注释研究》，侧重于自动校勘、自动编纂与自动注释研究，兼顾古籍计算机整理与开发系统的构建与集成。作者常娥，博士，现就职于东南大学，硕士生导师。

《古籍计算机自动索引研究——以民国农业文献自动索引为例》，侧重于自动索引研究，并以民国农业文献自动索引为样本。作者王雅戈，博士、博士后，中国索引学会理事，现就职于常熟理工学院。

《古籍计算机全文数据库及内容挖掘研究——以〈方志物产·广东〉为例》，侧重于数据库内容挖掘研究，并以《方志物产·广东》之物产、引书等内容挖掘研究为样本。作者衡中青，博士，中国索引学会理事，现就职于佛山科学技术学院。

《古籍计算机信息门户自动构建与应用——以农史学科为例》，侧重于信息门户自动构建与应用，并以农史学科信息门户构建与应用为样本。作者刘竟，博士，现就职于江苏大学。

《农业历史文献数字化建设研究》，侧重于农史文献数字化实践——中国农业遗产信息平台建设，并介绍其实际应用。作者曹玲、薛春香，均为博士，分别就职于南京信息工程大学、南京理工大学。

本项目丛书的出版发行，可为正在有志于从事本领域研究和工作的人员提供一个可资借鉴的文本。我们期待本丛书能为中国从文化古国向文化大国、文化强国迈进尽绵薄之力。

目　录

1 绪 论

学科信息门户是组织学科网络资源的有效工具，为广大科研工作者查找学科网络资源提供了巨大便利。自其产生以来，引起了国内外学者和机构的普遍关注。本章将对学科信息门户的产生背景、定义、特点及国内外研究和建设现状进行总结和介绍，同时选取国内外著名学科信息门户进行对比，分析我国学科信息门户的不足且提出改进建议。

1.1 计算机信息门户概述与展望

1.1.1 产生背景

学科信息门户兴起于20世纪90年代，是伴随着因特网的发展而出现的一个新名词，是目前网络学科信息资源不断增加而质量却参差不齐的产物。

因特网作为当代信息存储与传播的主要媒介之一，是一个巨大的信息资源库。其内容包罗万象，涉及不同学科、不同领域的方方面面，包括了文本、图像、图形、音频、视频、动画等。因信息量的巨大且存取不受时空的限制，逐渐受到了科研人员的重视和欢迎，成为研究人员获取信息资源的重要途径。但是，由于网络信息的发布缺乏必要的过滤、质量控制与管理，导致各种学术信息、商

业信息和个人信息混在一起，信息质量鱼龙混杂，给用户利用信息带来了极大的不便。

搜索引擎的出现，在一定程度上解决了网络信息资源的有序化问题，并逐步发展成为因特网上最流行的信息组织工具。但现有搜索引擎的检索是基于关键词的全文检索，一次检索可能会有成千上万条检索结果，而检准率极低，用户需要从中挑选出满足需求的个别条目，检索者很难获得理想结果。随着科研人员对学术信息资源的需求日益加深和扩展，搜索引擎在查找学术信息方面显得力不从心。

因此，科研人员要想在互联网中查找某一学科的信息，往往需要耗费大量的时间、精力和费用，极大地限制了科研人员对网络学术资源的利用。

在网络信息资源飞速增长，但信息组织手段滞后，与特定用户的需求产生矛盾的背景下，图书情报界也不再仅仅局限于作为印刷资源的主要组织者，也加入网络信息组织的行列。他们将传统的分类、标引和组织的优势扩展到网络信息空间，开发出了一种新型的信息组织方式——学科信息门户（Subject-based Information Gateway，SBIG）。它的出现有效弥补了搜索引擎的不足，提高了网上资源的有序化程度。从20世纪90年代起，学科信息门户作为一种高质量网络信息资源的深层组织方式，已逐渐受到了国内外广大科研工作者的认可与欢迎，为他们检索网络学术资源节省了大量时间，提高了学习、工作、科研的效率。

1.1.2 定　义

学科信息门户，最初是在英国的电子图书馆计划（The Electronic Libraries Programme）中提出来的，其英文全称有“Subject Information Gateway”，“Information Gateway”及“Subject-based Infor-

mation Gateway”等多种提法；在我国，因对“Subject”和“Gateway”的理解和翻译不同，学科信息门户亦有“主题网关”、“学科信息网关”、“学科门户”及“主题门户”等多种中文称谓，通过对其研究论文的统计，“学科信息门户”名称的使用较为普遍。

目前，学科信息门户尚没有确切的定义，一些国内外机构、学者从不同的侧重面对之进行了阐述，较有代表性的有以下几种：

（1）澳大利亚学科信息门户联盟（The Australian Subject Gateway Forum，ASGF）将学科信息门户定义为一种用于获取高质量的、经过评价的资源的网络机制，其资源的评价标准为是否支持特定学科主题的研究或者学习[1]。

（2）DESIRE 项目提出的定义为：“学科信息门户是提供网络资源的可检索和浏览目录的在线服务和网站，主要关注相关学科领域的学术信息。”同时，在其学科信息门户手册（*DESIRE Information Gateways Handbook*）中对 SBIG 的特征进行了归纳：“学科信息门户是高质量和可控信息服务资源，具有以下特征，提供网上大量网站或文献的链接服务；根据特定的质量和范围标准运用人的智力劳动过程选择资源；依靠人的智力完成内容描述；依靠人的智力构建结构（不包括完全无组织的链接表）；至少部分是人工为每个资源创建（书目）元数据。”[2]

（3）学科信息门户概念最早的提出者 Traugott Koch 认为：“学科信息门户是支持系统化资源发现的因特网服务，通过因特网提供对资源（文献、对象、网站或服务）的链接。该服务建立在资源描述的基础之上，可以通过主题结构浏览和访问资源是其重要特征。”同时，还进一步将学科信息门户分为一般的学科信息门户和质量受控的学科信息门户：描述较少、主题结构肤浅的链接列表，为一般的学科信息门户；有着丰富的资源描述、遵循一定标准和深层次的分类结构、有着高标准的质量受控的学科服务，为质量受控的学科

信息门户[3][4]。

（4）国内学者张晓林教授认为，学科信息门户“致力于将特定学科领域的信息资源、工具与服务集成到一个整体中，为用户提供一个方便的信息检索和服务入口”[5]。

（5）王楠、吴新年、祝忠明在文献中提出：“学科信息门户是针对特定学科或主题领域，按照一定的资源选择和评价标准、规范的资源描述和知识组织体系，对具有一定学术价值的网络资源进行搜集、选择、描述和组织，并提供浏览、检索、导航等增值服务的专门性信息门户。”[6]

以上观点从不同角度分析了学科信息门户的概念。综合以上观点，笔者认为，可将学科信息门户描述为：学科信息门户是一种在线网络信息集成机制和服务，它针对特定学科或主题领域，制定相关的资源选择和评价标准，对具有一定学术价值的、高质量的网络资源进行搜集和选择，利用规范的资源描述和知识组织体系对其进行描述和组织，将网络中特定学科领域的信息资源、工具与服务集成到一个整体中，并提供浏览、检索、个性化等增值服务，旨在为用户提供一个方便的网络学术信息检索和服务入口，以满足用户科研和教育的信息需求。

1.1.3 相关概念分析

门户、网络资源导航库、数字图书馆等都是近几年发展起来的网络资源组织的新形式，这些概念与学科信息门户既有密切联系，又有一定区别，辨别这些概念之间的关系可以更好的理解、认识学科信息门户。

（1）门户。又称信息门户，在英文可用 gateway、portal、information portal 表示，是指一定范围内以某种方式提供对其他资源的获取的 Web 站点或 Web 服务，通过自定义类目或面向主题分级列表