
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 ZIDONG SUOYIN YANJIU
YI MINGUO NONGYE WENXIAN
ZIDONG SUOYIN WEI LI

# 古籍计算机自动索引研究

## ——以民国农业文献自动索引为例

王雅戈◎著

安徽师范大学出版社

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系列

丛书主编◇侯汉清

GUJI JISUANJI ZIDONG SUOYIN YANJIU
YI MINGUO NONGYE WENXIAN
ZIDONG SUOYIN WEI LI

# 古籍计算机自动索引研究

## ——以民国农业文献自动索引为例

王雅戈◎著

安徽师范大学出版社

责任编辑：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**

古籍计算机自动索引研究：以民国农业文献自动索引为例/王雅戈著．—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 978-7-5676-1000-2
Ⅰ.①古…　Ⅱ.①王…　Ⅲ.①农业技术—古籍—索引—编制—自动化—研究　Ⅳ.①G353.21-39

中国版本图书馆 CIP 数据核字（2013）第 238943 号

古籍计算机自动索引研究
——以民国农业文献自动索引为例
王雅戈　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：13.75
字　数：186 千
书　号：ISBN 978-7-5676-1000-2
定　价：32.00 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　中国古籍索引史研究

索引是治学的利器，是传统方式下最有力的信息组织基础工具之一。索引具有揭示文献主题内容、指示文献来源的功能，还有校雠学和语言学方面的功能，可以为文献检索、阅读和利用提供极大便利。索引在我国曾有辉煌的历史，梳理索引学术史，特别是对民国时期农业文献索引史进行回顾，对于数字化条件下的农史文献信息组织能起到一定的参考作用。从清代著名史学家章学诚到近代学者梁启超、胡适、林语堂等，都对索引的作用有深刻的认识，提出了许多有益的见解，对索引学的启蒙与发展起了很大的推动作用。特别需要提到的是，曾位居民国政府代总理、外交总长等要职的蔡廷干，编纂了我国第一部语词索引《老解老》。历史学家洪业领导下的哈佛燕京学社引得编纂处，编纂了大量古籍索引。洪业本人还撰写了我国近现代索引学研究的开山之作《引得说》。欧美及日本等国学者，也研究和编纂了许多中国古籍索引。总之，在近代学术园地里，众多中外学者积极参与索引事业理论探讨和实践研究，成就了中国近代索引事业的繁荣与辉煌。

## 1.1　中国古籍索引概述

### 1.1.1　古籍索引基础知识

#### 1.1.1.1　概　念

古籍索引是揭示古籍内容的一种特定形式，但不同于古籍书

目。古籍书目所著录的是古籍书名、卷数、著者、时代、版本等款目，将其按一定次序编排，以揭示与报导古籍文献的外形特征和内容梗概，供人们查阅的工具；古籍索引则是将古籍中的有关事物名称，或篇名、或字句、或词语、或人名、或地名、或内容主题等，分别摘录勾引、注明出处、页码行数，并按一定排检方法编辑而成的供人们查寻有关古籍文献内容的工具。在揭示古籍文献内容方面，古籍索引比古籍书目更有深度，更为详尽而具体。

中国古籍索引大体上分为“中国古籍原文索引”和“中国古籍研究论著分类索引”两大类，关于这两类索引的编制在不同的历史时期以及不同的地域则各有侧重。其中东方学者更加注重“中国古籍原文索引”。

#### 1.1.1.2 简　史

就中国来说，自明代万历三年（1575）张士佩首次编制索引——《洪武正韵玉键》以来，索引编制为历代文人雅士所重，编制不绝如缕，各种古籍索引成果亦蔚然可观。自明以降，《两汉书姓名韵》（明末傅山编制）、《本草万方针线》（清代蔡烈先编制）等继张氏之踵，于索引一道亦深加探究。迨及民国，国内学者纷纷采用国外索引编制技术，索引数量激增，其实用性有显著增强。影响最大、泽溉至广者，当推哈佛燕京学社引得编纂处和中法汉学研究所通检组所编制的一系列引得、通检。1949 年以来，古籍索引工作取得了很大的成就。据统计，1949—2006 年，全国各出版社共出版古籍索引类图书 135 种，索引编制工作成就显著。

然而古籍索引编制并非仅为我国学者所重视，日本、欧美等亦成绩斐然，其中尤以日本为最。1975 年美国人麦克马伦编制的《中国典籍索引》由中国资料研究服务中心于 1975 年印行，标志着美国人首次编制的中国古籍索引作品正式面世。日本人长期以来对中国古籍索引较为关注。据陈东辉先生统计，日本所出版的中国古籍

索引约占全世界所出版的中国全部古籍索引总量的80%，仅唐代古籍索引就达38种，日本学者对中国古籍索引的关注程度之高由此可知。日本所编古籍索引涉及中国传统的各个领域，经、史、子、集四部均有相应成果面世。

#### 1.1.1.3　类　型

古籍索引的种类远较现代文本的索引来得丰富多样，不仅可以按照常规的分类方法将索引划分为内容索引与篇目索引（按索引功用分），印刷品、缩微品以及电子产品索引（按载体形态分），等等，还有自己独特的类型。如：

第一，按索引源的数量分类，古籍索引分为专书索引和群籍索引，面临不同的对象时，其称呼也略有差异。在史学领域，古籍索引可以分为群史索引与专史索引，在地方志中则分为群志索引与专志索引，另外在诗文索引中，通常称为专集索引与总集索引。

第二，按索引的标目分类，除了常见的主题索引、著者索引、地名索引、题名索引、全文索引等外，古籍索引亦有自身特质，比如属于全文索引的逐字索引、词汇索引、句子索引、引书索引等，这些索引往往在古籍索引中具有较高的比重。在主题索引中，古籍索引的品种亦最多，比如植物索引、动物索引、物产索引、农具索引、史实索引、图谱索引等。

第三，按索引的编排和组织方式分类，一般可以分为分类索引、字顺索引以及分类—字顺索引三类。其中字顺索引种类尤其繁多。古籍字顺索引大多数采用了“部首＋笔画”、中国字庋撷法、四角号码法等形式。将索引款目按某种分类体系或按某种分类标记排列的索引，通常称为分类索引。中国历史典籍种类浩繁，且采用的分类体系各异，因而其分类索引也多有不同。其中四库法索引为最多。当然，并不是所有的古籍分类索引均采用四库法。

第四，按索引的资料来源分类，因为所依据的资料不同，索引

的种类也不尽相同。具体有以下七个大类：

经典索引，以各种典籍为标引来源而制作的索引。佛教经典索引、道教经典索引与儒家经典索引皆是此类。

诗文索引，以古代诗词文赋为主要索引对象。诗文集一般分为总集与别集，与此相对应，其索引也分为总集索引与别集索引。

词曲索引，词曲中的人名、剧名、曲牌名、语词及典故，均可作为索引标目，编制出不同风格的索引。

方志索引，学术界将方志分为新方志、旧方志两大类。据衡中青博士研究，旧方志索引始于清代的《同治鄞县志》，其总目录卷26至卷44详列人物姓名共54页，分别按卷、类、时代先后排列，已经具备了现代人名索引的一些要素。20世纪30年代以来，旧方志编制索引逐渐活跃。

类书索引，以类书为标引来源而编制的索引。类书篇幅浩繁，内容包罗万象，“贯穿古今，汇合经史，天文地理，皆有图记。下至山川草木，百工制造，海西秘法，靡不备具。洵为典籍之大观”。为之编制索引难度较大，需要专门的机构与组织来实施。编制《古今图书集成索引》电子版前后历时10年，是一次成功的尝试。

丛书索引，以丛书作为标引来源而编制的索引。在中国古籍中，丛书数量甚多。此类古籍，学者非常重视，编制了各种索引。例如，《四库系列丛书目录·索引》就是针对四库系列七种丛刊的资料而编写的索引，规模空前。另外几种使用最为广泛的索引为《丛书子目索引》、《丛书子目书名索引》、《匹库荟要目录索引》、《四部备要索引》（诸家骏编，中国台北中华书局1971年版）、《百部丛书集成人名索引》（中国台北艺文印书馆1971年版）等。

诗话索引，以诗话中的人名、篇名、名句、诗体名、流派名、术语、事项编为标目的索引，将大大有助于古代诗歌理论的研究和诗歌史料的钩沉。《索引本何氏历代诗话》（马汉茂编，中国台北美