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
QUANWEN SHUJUKU JI
NEIRONG WAJUE YANJIU
YI FANGZHI WUCHAN GUANGDONG WEI LI

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

衡中青◎著

安徽师范大学出版社

国家出版基金项目
NATIONAL PUBLICATION FOUNDATION

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
QUANWEN SHUJUKU JI
NEIRONG WAJUE YANJIU
YI FANGZHI WUCHAN GUANGDONG WEI LI

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

衡中青◎著

安徽师范大学出版社

责任编辑：郭行洲　　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

古籍计算机全文数据库及内容挖掘研究：以《方志物产·广东》为例/衡中青著.
—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 978-7-5676-0999-0
Ⅰ.①古… Ⅱ.①衡… Ⅲ.①古籍—全文数据库—数据采集—研究 Ⅳ.①G256.1

中国版本图书馆 CIP 数据核字（2013）第 238918 号

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
衡中青　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
字　　数：151 千
书　　号：ISBN 978-7-5676-0999-0
定　　价：24.00 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

1.1 地方志目录学整理概况

中国地方志以起源早、持续久、类型全、数量多而享誉于世界。据《中国地方志联合目录》[1]的统计，仅保存至今的宋代至民国时期的方志就有8 264种，11万余卷，占中国古籍的十分之一左右。但实际不止于此，还有遗漏，再加上未计入的山水寺院志，旧志当应在1万种左右。新中国成立后，倡修新志，所编修的各种志书也在1万种左右。这样，旧志、新志的总量在2万种以上[2]。所以，中国地方志可谓卷帙浩繁，无疑是地方文献的大宗。

地方志记载的是某一地区自然、历史、地理、社会、经济和文化等各个方面的情况和资料，史料范围十分广泛，它所汇集的某一地方的基本知识和系统资料宏富广阔，是一种地方性百科全书。因此，地方志既具有丰富、坚实的史料基础，又具备取之不尽、足资参证的资料，价值极其巨大。然而，地方志所载的资料庞杂，质量参差不齐，收藏遍及世界各地，极具分散性。这些不利因素无疑给我们利用、查阅地方志造成极大的不便。有鉴于此，有必要开展地方志目录学研究，编制地方志目录，报道地方志馆藏，揭示地方志内容，以便更好地读志、用志。

1.2 方志目录学概念

我们知道，地方志史料价值巨大。目前，中国地方志这份珍贵遗产，受到各行各业重视，得到广泛的利用。然而，地方志利用起来十分困难，解决这个困难最有效的办法之一，是编纂地方志目录。地方志目录是打开地方志大门与了解地方志内容的钥匙。

方志目录学是方志学与目录学的边缘学科。方志目录是一种专科目录，是研究、整理和利用地方志的指南。它是以目录学为指导，以方志目录的编纂与社会需要为研究对象，探究其历史与源流、编辑原则与方法以及发展趋势的学科。它是从方志读者群的需要出发，揭示与报道方志学科文献状况的学科。

方志目录学这一术语，在许多前辈的论著中曾多次被提及，但鲜有从理论上作专题性和系统性的阐述，直至20世纪末和21世纪初才有专文讨论这一问题[3-5]。但是，所讨论的对于作为一门学科的方志目录学来说，不够全面、系统和深入，且多为对方志目录成果的评述。因此，笔者认为，到目前为止，方志目录学作为一门学科体系还没有建立起来，还没有从理论高度上对方志目录实践作出全面总结和深入研究。然而，当前人们对方志文献的利用越来越频繁，对方志文献的情报需求越来越深入，因此，全面深入开展方志目录学研究，具有重要的现实意义。

1.3 方志目录源流

方志著录于书目始于南朝。刘宋时期王俭的《七志》载："七曰图谱志，纪地域及图书。"梁阮孝绪著《七录·传录》，设"土地部"收地记、地志，首发端绪。到了唐代魏徵等撰《隋书·经

籍志》时，其“地理部”首开将方志纳于正史艺文志的传统。《隋书·经籍志》的“地理部”及宋代的《直斋书录解题》“史部地理类”，统辖地志、地记、图经、图志，奠定了封建时代方志目录学的体制——综合目录、综合提要的基础。此后，方志不断发展，数量逐渐增多，在目录学发展中的地位也不断提高。唐宋以降，有个别学者在整理方志时另创一格，著录成独立的部类，如：明代朱睦㮮整理万卷堂藏书，编《万卷堂艺文志》，将“地志”移出“地理类”，列史部之十一；明代张萱校理内阁藏书，于经、史、子、集外另立“志乘”部，收录正统至万历间庋藏志书；清道光中期常熟钱曾编的《述古堂书目》卷三有“地志”类。但就绝大部分目录整理而言，均未脱离传统目录学的范畴。清末民初，随着西方史学、地理学、目录学的传入，图书整理逐步冲破了传统目录学之桎梏，开拓了新研究领域，方志整理也进入了一个新历史时期。1912 年，我国近代第一部影响很大的方志工具书——《清学部图书馆方志目》终于问世。从此以后，不同的单位和个人陆续编制了数量众多、类型多样的方志书目，出现了“联合目录”这一前所未有的类型。

据上所述，清末以前的方志在目录学范畴上还处于萌芽时期，还没有明确的方志专门目录，或者说专门目录没有出现。清末以前的方志目录体制是综合目录、综合提要，这也是封建时代方志目录最主要的类型。

1.4 方志目录类型

有关方志目录的类型，目前还没有统一的划分标准[6]。综观古今方志目录的发展演变，其类型还是复杂多样的。按著录的范围划分，有馆藏的、综合的；按编制的形式分，有书本式、附录式（附

录于图书之后）、期刊式（发表于杂志）、卡片式；按反映的内容分，有宋元方志目、明代方志目、通志目、乡土志目等。

综观方志目录的发展史，笔者提出一种划分方法，即按照历史演变形式，方志目录可分为综合目录（综目）、专门目录（专目）和修志目录三种，其中专门目录又可分为馆藏专目、联合目录两种。

（1）综合目录。综合目录是指全面著录图书庋藏单位各类图书的目录，所录的图书涉及多个知识门类。这种目录类型中，方志目录与其他知识门类的目录一样，是综合目录的有机组成部分之一。综合目录的优点是便于按学科进行查阅，缺点是著录方志的数量有限。

史志艺文志载录方志，始于《隋书·经籍志》，其后的《唐书·经籍志》、《新唐书·艺文志》、《宋史·艺文志》、《新元史·艺文志》、《明史·艺文志》亦沿此例。

方志目录独成一目者，始于《通志·艺文略》，其地理类有郡邑之属、图经之属，这是方志专类之滥觞[7]。其后，《国史经籍志》地理类有图经一门，《万卷堂艺文志》地理类有方志一门，《澹生堂藏书志》图书类有省会通志、郡邑志二门。《清史稿·艺文志》地理类有都会郡县之属，所载大都是官修之书；黄虞稷所撰《千顷堂书目》，其目虽为地理类，实际上所著录的基本上为方志，多达1 600种。《四库全书总目》史部地理类分为九目：总志、都会郡县、河渠、边防、山川、古迹、杂记、游记、外纪。其中，总志、都会郡县二目专收方志图书。总之，方志综目是我国封建时代方志传统目录的主要形式。

（2）单一收藏单位之馆藏专目，常简称为馆藏专目。顾名思义，这是某个收藏单位所编的反映本馆方志收藏情况的专门目录，这里的收藏单位只能是一个，多于一个则属于联合目录的范畴。现存首部馆藏方志专目，为中华民国2年（1913年）缪荃孙编制的

《清学部图书馆方志目》。馆藏方志大发展是20世纪20—30年代的事。馆藏方志专目以单个收藏单位为著录范围，数量有限，加上分散全国各地，甚至全世界，因此读者难以获取，更不知有哪些收藏单位编有方志专目。

（3）多个收藏单位之联合目录。方志联合目录是著录多个收藏单位的方志目录联合体，揭示多个收藏单位的收藏方志情况，它可以著录两个收藏单位的馆藏、某一地区或几个地区的馆藏，甚至全国性的、多个国家的馆藏。联合目录较单一馆藏专目有方便快捷之利，查找一种联合目录，即可知何种方志存何馆、何馆有何种方志。馆藏方志联合目录创始于20世纪30年代，以朱士嘉《中国地方志综录》为代表。目前质量最高、使用最广泛的馆藏方志联合目录为《中国地方志联合目录》，是在《中国地方志综录》的基础上不断增订而成。馆藏方志联合目录的特点是，以馆际协作的形式，体现方志收藏地点，揭示方志的分布状况，以实现馆际间的方志资源共享。

（4）修志目录。这是一种特殊的目录，它全面反映某一时代或某一区域的方志编纂、刊印情况，是一定历史时期和一定地域方志编修的历史记录。修志目录分为两种类型：一是在著录国家或地区的图书出版状况时记录地方志的编修出版情况的修志综目，如《明史·艺文志》、《全国新书目》等；二是修志专目，如《中国新编地方志目录》、《浙江新编地方志目录》等。修志目录属于登记类目录，特点是反映面广而全，反映的信息快而准。

1.5 近百年方志目录述评

1.5.1 中华民国时期方志的目录学成就

（1）馆藏方志专门目录兴起[8]。方志有专门目录，据《隋