<<古籍计算机断句标点与分词标引研究>>

图书基本信息

书名：<<古籍计算机断句标点与分词标引研究>>

13位ISBN编号：9787811416749

10位ISBN编号：7811416743

出版时间：2012-2

出版时间：安徽师范大学出版社

作者：黄建年

页数：148

内容概要

《古籍计算机断句标点与分词标引研究》以古籍中的农业文献及农史信息资源为研究对象，利用计算机技术及现代情报技术进行整理与开发，但各册研究重点并非完全一致，或标点与分词，或编纂与校勘，或知识组织，或内容挖掘，或索引编制，或关注信息门户，或研究古籍数字化技术。虽各有分工、各有侧重，但却互相补充、紧密联系。

作者简介

黄建年，男，1966年生，研究馆员，1990年毕业于北京大学图书馆学专业本科，获得文学学士学位，2004年、2009年分别于南京大学、南京农业大学获得管理学硕士、理学博士学位。
先后担任图书馆办公室主任、图书馆馆长助理、物资设备处副处长等职。
曾兼职担任镇江市图书馆学会学术委员会副主任，现为江苏省黄氏文化研究会(筹)常务理事。
主持或者参与国家社科基金项目、教育部人文社会科学基金项目、江苏省教育厅高校社科基金项目等10余项，在《中国图书馆学报》、《情报学报》等各类刊物上发表学术论文60余篇，出版专著2部，开发软件2种。
主要研究方向：古籍整理、数字图书馆、信息组织。

侯汉清，南京农业大学信息科技学院教授、博导，中国索引学会副理事长。
主要专著有《主题法导论》(1991年)、《索引技术和索引标准》(1997年)、《当代分类法主题法索引法研究》(1997年)、《文献分类法主题法导论》(1999年)、《图书馆学五定律》(译著，1984年)、《引文索引法的理论及其应用》(译著，2003年)、《情报检索语言与智能信息处理丛书》(主编，2009年)等。
主持或参与主持国家级项目6项，主编或参与主编《中国分类主题词表》、《社会科学检索词表》等国内主要词表和分类表多部。
研究方向：信息组织、信息检索、信息智能处理。

书籍目录

章节摘录

版权页：插图：索引编制自动化主要集中在逐字索引，而对词的索引研究不多，所以本论文的研究重点在索引词汇的识别，通过自动识别索引词汇，然后实现索引的自动编制，编制出真正基于语词的古籍索引。
3.2 分词标引理论与关键技术研究3.2.1 分词词典研究分词词典是基于词典分词的汉语自动分词系统的一个组成部分，也是影响系统性能的重要因素之一。
基于词典的自动分词系统所需的各类信息基本上从分词词典中获取。
考核分词词典质量主要有两个指标：①词典的内容，即词典中收录词汇的数量与质量，它对分词精度有着很大的影响。
一个好的词典要具备通用性好、覆盖率高的优点。
②词典的组织形式。
系统在进行分词及标注时需要频繁地查询词典，词典的查询速度直接影响到分词系统的速度，因而必须有效地组织词典，从而提高系统的整体性能。
建立词典有两种方法：①建立静态词库，这是一种简单直接的方法。
静态词库以国家技术监督局1993年发布的GB／T13715-92《信息处理用现代汉语分词规范》为依据，其特征是针对信息处理的基本需要、以人为本、考虑词的常用性，整个词表分成词库、带字母词库、专名库、常用接续库、成语库、俗语库以及单字词库等7个分词库。
该规范具有较强通用性及覆盖能力，对推动汉语自动分词研究的发展，起到了积极作用。
但该方法的不足之处在于字典所能包含的单词有限，对于特定领域的某些单词无法包含。
并且对于某个特定领域的应用来说，实际需要的单词要少得多，大而全的字典反而影响分词的效率和准确率。
②建立动态词库，动态词库也称为智能词库，使用统计方法通过对大量的语料文本进行处理来建立词典。
智能词典的基本思想是：先用无词典法按照一定算法对分词文本进行特征提取，提取出中频词与高频词两类，按照一定的算法决定高频词是否为新词，若有新词则添加到临时词典，然后按照机械分词算法进行分词。
智能词典定时地对临时词典进行处理，按照一定的算法提取特征词，将其放人词典。
该方法针对特定领域，词典的规模相比于通用词典要小得多，其分词的效率比第一种方法要高。
更重要的是，这种方法统计大量的语料文本，能包括本领域的几乎所有单词，其分词的准确率也比前一种方法高。

编辑推荐

版权说明