“十三五”国家重点出版物出版规划项目

·藏文信息处理技术·

བོད་ཡིག་ཡིག་རྟགས་རང་འགྱེད།

—— Visual C++ཡིས་སྒྲུབ་པ།

现代藏文字符自动分析

——Visual C++实现

高定国◎著

西南交通大学出版社

“十三五”国家重点出版物出版规划项目

藏文信息处理技术

བོད་ཡིག་ཡིག་རྟགས་རང་འགྱེད།

—— Visual C++ཡིས་སྒྲུབ་པ།

现代藏文字符自动分析

—— Visual C++实现

高定国　著

西南交通大学出版社

·成都·

图书在版编目（CIP）数据

现代藏文字符自动分析：Visual C++实现 / 高定国著. —成都：西南交通大学出版社，2022.3
（藏文信息处理技术）
“十三五”国家重点出版物出版规划项目
ISBN 978-7-5643-7904-9

Ⅰ. ①现… Ⅱ. ①高… Ⅲ. ①C语言－程序设计－应用－藏文－语言信息处理学 Ⅳ. ①TP312.8②TP391

中国版本图书馆CIP数据核字（2020）第255295号

“十三五”国家重点出版物出版规划项目
（藏文信息处理技术）
Xiandai Zangwen Zifu Zidong Fenxi
——Visual C++ Shixian
现代藏文字符自动分析
——Visual C++实现

高定国　著

出 版 人　王建琼
责任编辑　穆　丰
封面设计　墨创文化

出版发行　西南交通大学出版社
（四川省成都市二环路北一段111号
西南交通大学创新大厦21楼）
邮政编码　610031
发行部电话　028-87600564　028-87600533
网址　http://www.xnjdcbs.com
印刷　四川森林印务有限责任公司

成品尺寸　210 mm×285 mm
印张　18.25
字数　528千
版次　2022年3月第1版
印次　2022年3月第1次
定价　88.00元
书号　ISBN 978-7-5643-7904-9

课件咨询电话：028-87600533
图书如有印装质量问题　本社负责退换

前 言
preface

教育部高等学校计算机科学与技术教学指导委员会发布的《高等学校计算机科学与技术专业人才专业能力构成与培养》中对计算机科学与技术专业的基本能力要求确定为：计算思维能力、算法设计与分析能力、程序设计与实现能力和系统能力。对计算机专业的学生来说，算法设计与分析、程序设计与实现是最主要和最重要的能力。本书依据本人多年讲授“算法设计与分析”和“藏文信息处理原理”课程的经验，以藏文信息处理的知识为内容，并结合算法设计与分析的相关方法编写而成。本书不仅适用于藏文基础较好、已对计算机初步入门的学生学会藏文在计算机上如何处理，而且也适用于计算机基础较好、拟从事藏文信息处理的人员掌握藏文字符处理的基本过程，旨在跨越藏文字符分析的设想与计算机实现的鸿沟。

本书在编写中努力体现了以下特点：

案例贯穿，规范设计。本书以案例式进行编排，每一章就是一个案例。按照算法设计与分析的思想，每章分别由问题描述、问题分析、算法设计、程序实现、运行结果和算法分析等步骤组成。该方法符合人们发现问题、分析问题、解决问题的思维习惯，也符合软件工程的问题分析、软件设计、软件实现与软件测试的软件设计思想。

循序渐进，层层深入。本书的内容从藏文字符的输入输出开始，逐步到全藏字符的生成，现代藏字的构件识别，基于不同算法的藏文字符排序、查找和字符属性统计。编程使用的软件从控制台应用程序到基于对话框的可视化程序。编程文件从“单一文件”逐步到“头文件”“源文件”分开的“两个文件”，再到多个类的“多文件”。

案例详尽，步步指引。所有的案例在“程序实现”中都详尽地给出了每一步的实现方法和源代码。读者可以按此程序的实现，步步复现程序。

拓展知识，轻松实现。每个案例的“理论依据”并没有局限于本案例算法设计的理论基础，而是拓展了实现本案例关键的编程知识，使得读者能够轻松用计算机实现各案例。

全书分为 4 篇，共 16 章。藏文字符处理基础篇主要介绍算法概述、藏文字符的输入输出、全藏字的生成等藏文字符处理基础；藏文字符排序篇以现代藏字构件识别为基础分别用插入排序、归并排序、堆排序、快速排序实现了藏文字符的排序；藏文字符查找篇运用查找算法实现了藏文编码转换、藏文的拉丁转写、藏文数字编码方案和藏汉电子词典；藏文字符统计篇实现了全藏字集字符构件的静态统计、藏文多文本中

构件的动态统计、基于动态顺序存储的藏文音节动态统计、基于哈希表的藏文音节动态统计等藏文字符的统计。

本书是在国家自然科学基金项目“敦煌藏文文献文本识别方法的研究”（62166038）、教育部 2019 年第一批产学合作协同育人项目“改革教学内容和课程体系，提升应用型人才的开发能力”（201901077046）、2019 年度国家级一流本科“计算机科学与技术”和“藏文信息处理原理”一流本科课程建设的资助下取得的成果之一。作者在编写过程中得到了西藏大学信息科学技术学院领导、同事的支持和帮助，书中部分案例是在研究生课程设计的基础上改编而来的，在此一并表示感谢！

本书可以作为高等院校藏文信息技术、计算机科学与技术、电子信息技术等相关专业的高年级本科生或研究生的教材或参考书，也可以作为从事藏文信息处理、自然语言处理、藏语计算语言学、数据挖掘和人工智能研究相关人员的参考书。

编著人员水平有限，加之时间仓促，参考资料缺乏。书中难免有疏漏与不足之处，恳请广大读者批评指正。

高定国

2022 年 2 月

目 录
contents

第 1 篇　藏文字符处理基础

第 2 篇　藏文字符排序

第 3 篇 藏文字符查找

第 4 篇 藏文字符统计

第1篇　藏文字符处理基础

第 1 章　算法概述

一般来说，用计算机解决一个具体问题一般需要经过以下几个步骤：首先要从具体问题抽象出一个适当的数学模型，其次设计一个解此数学模型的算法，然后编出程序，最后对程序进行测试、调试直至得到最终解答[①]。可以看出，算法设计是计算机解决一个问题的核心。那么什么是算法呢？

1.1　算法的概念

算法（Algorithm）[②]是指对解题方案准确而完整的描述，是一系列解决问题的清晰指令，代表着用系统的方法描述解决问题的策略机制。也就是说，算法能够实现对一定规范的输入，在有限时间内获得所要求的输出。算法中的指令描述的是一个计算，当其运行时能从一个初始状态和（可能为空的）初始输入开始，经过一系列有限而清晰定义的状态，最终产生输出并停止于一个终态。

1.2　算法的特征

一个算法应该具有以下 5 个重要的特征：

1. 有穷性（Finiteness）

算法的有穷性是指算法必须能在执行有限个步骤之后终止。

2. 确切性（Definiteness）

算法的每一步骤必须有确切的定义。

3. 输入项（Input）

一个算法有零至多个输入，以刻画运算对象的初始情况。所谓零个输入是指算法本身定出了初始条件。

4. 输出项（Output）

一个算法有一个或多个输出，以反映对输入数据加工后的结果。没有输出的算法是毫无意义的。

5. 可行性（Effectiveness）

算法中执行的任何计算步骤都是可以被分解为基本的可执行的操作步骤，即每个计算步骤都可以在有限时间内完成（也称之为有效性）。

---

① 严蔚敏，吴伟民. 数据结构（C 语言版）[M]. 北京：清华大学出版社，2017.

② 钟志永，姚珺. 大学计算机应用基础[M]. 重庆：重庆大学出版社，2012.

1.3 算法分析

对于一个实际问题，通常可以提出若干个算法来解决。如何从这些可行的算法中找出最有效的算法呢？或者有了一个解决实际问题的算法后，如何来评价它的好坏呢？这些问题都需要通过算法分析来确定。评价算法性能的标准主要从算法执行时间和占用存储空间两个方面进行考虑，即通过分析算法执行所需要的时间和存储空间来判断一个算法的优劣[①]。算法分析就是对一个算法需要多少计算时间和存储空间做定量的分析[②]。

分析算法可以预测这一算法能在什么样的环境中有效的运行，能对解决同一问题的不同算法的有效性做出比较[③]。

1.3.1 时间复杂度

一个程序的时间复杂度是指程序运行从开始到结束所需要的时间。

1. 影响因素

一个算法是由控制结构（顺序、分支和循环3种）和原操作（固定数据类型的操作）构成的，其执行时间取决于两者的综合效果。为了便于比较同一问题的不同算法，通常的做法是：从算法中选取一种对于所研究的问题来说属于基本运算的原操作，以该原操作重复执行的次数作为算法的时间度量。一般情况下，算法中原操作重复执行次数是规模 n（即算法处理的数据量）的某个函数 $T(n)$。很多时候要精确地计算 $T(n)$是困难的，通过引入渐进时间复杂度在数量上估计一个算法的执行时间，也能够达到分析算法的目的[④]。

2. 计算方法

计算时间复杂度的时候，主要考虑算法中最高阶项的开销，只要找出算法中最高阶的复杂度，就可以忽略低阶和常数的复杂度。

这里引入数学符号“O”来估算算法时间复杂度，渐进时间复杂度的表示方法为：

$$F(n)=O(g(n))$$

其定义为，若 $F(n)$ 和 $g(n)$ 是定义在正整数集合上的两个函数，则 $F(n)=O(g(n))$ 表示存在正的常数 c 和 n_0，使得当 $n\geqslant n_0$ 时，都满足 $0\leqslant F(n)\leqslant cg(n)$。换句话说，就是这两个函数的整型自变量 n 趋于无穷大时，两者的比值是一个不等于0的常数。

当要计算某个算法的时间复杂度 $F(n)$ 时，可以找一个更简单的、阶数相同的时间复杂度 $g(n)$ 来等同地计算，这里的 $g(n)$ 是替代函数，它具有和原算法一样更高阶复杂度。例如，一个程序的实际执行时间为 $T(n)=3n^3+43n^2+5\,342$，则 $T(n)=O(n^3)$。使用 O 记号表示的算法的时间复杂度，称为算法的渐进时间复杂度。

通常用 $O(1)$表示常数计算时间。常见的渐进时间复杂度之间的关系如下：

$$O(1)<O(\log_2 n)<O(n)<O(n\log_2 n)<O(n^2)<O(n^3)<O(2^n)$$

① 程玉胜. 数据结构与算法C语言版[M]. 北京：中国科学技术大学出版社，2015：5-10.
② 程玉胜. 数据结构与算法（C语言版）[M]. 北京：中国科学技术大学出版社，2015.
③ 李长云，蒋鸿，刘强. 大学计算机[M]. 北京：北京航空航天大学出版社，2013：171-175.
④ 陈承欢. 数据结构分析与应用实用教程[M]. 北京：清华大学出版社，2015：9-11.