

计算机中文信息处理

上册

赵珀璋

徐力

编著

计算机 中文信息处理

上册

宇航出版社

内 容 简 介

本书全面系统地论述了计算机的中文信息处理和中文信息的计算机处理有关基础知识、基本原理、设计要领和使用方法。全书共分十四章,先后讨论了现代汉语基础;汉字输入编码方法;中文字形库;中西文兼容的检索技术、操作系统、程序设计语言、数据库、通信网络、信息处理系统;汉字识别和汉语语音识别;自然语言理解;少数民族语言文字处理;中文信息处理的有关国家标准。上册为前八章。

这是第一本作为我国研究生教材的中文信息处理技术的书籍,不仅可作为高等院校有关专业的教科书,也可供从事计算机科学、计算机工程、计算机应用和中文信息处理系统研制、生产、开发和使用部门的科技、管理人员参考。对提高和普及我国各地区,包括少数民族地区计算机的广泛应用,对文字改革及语言研究将会起到积极促进作用。而对世界各国研究中文信息处理技术,也将起到一定的推动作用。

计算机中文信息处理(Jisuanji Zhongwen Xinxi Chuli)

上 册(Shangce)

编著:赵珀璋 (Zhao Po-zhang)

徐 力 (Xu Li)

责任编辑:廖寿琪 (Liao Shou-qi)

宇航出版社出版 新华书店北京发行所发行
铁道出版社激光照排车间用华光型机进行照排
各地新华书店经售 铁道出版社印刷厂印刷

787×1092 1/16开本 印张27.5 字数:692千字
1987年12月第1版 第1次印刷 印数:1-8000册

统一书号:15244·0107 定价:5.80元

ISBN7-80034-000-7/TP·001

我们的祖先创造了中国文字,发明了造纸工艺和活字印刷术,才使我们具有今天灿烂的文化和文明。

电子计算机的发明,对于人类科学文化生活,具有划时代的意义。电子计算机已经历了四个发展阶段,现正向新一代即智能时代迈进。新一代计算机智能表现之一,就是人机对话录用自然语言或近似自然语言,在我国就是采用中文。

电子计算机诞生在西方,其软硬件系统适应于西方拼音文字。西方拼音文字字符数量少,结构简单。而中文,尤其是汉字,数量多,结构复杂。如何利用计算机来处理中文信息呢?

目前,计算机处理中文信息有两方面问题:一方面是如何改造现有计算机,使其适应中文信息的特点,即“计算机的中文信息处理”问题;另一方面是如何改革中文,使其适应计算机处理,即“中文信息的计算机处理”问题。要彻底解决这些问题,仅靠计算机科学家不行;仅靠语言文字学家也不行;还需要数学家、物理学家、电子学家、生物学家、心理学家、历史学家、逻辑学家等诸学科专家。

汉字的三要素是形、音、义。汉字的发音特点是单音节。汉字是二维图形。汉字的字数多、字体多、字号多。这些都与现行计算机代码体系,软硬件结构相矛盾。

计算机的字符顺序处理与中文信息的汉字横向从左到右从上到下,竖向从上到下从右到左排列;蒙文竖向从上到下从左到右排列;维吾尔文(哈萨克文)横向从右到左从上到下的排列等特点相抵触;

从语言文字学角度来看,需要专门的中文信息处理系统。从世界科技文化共性来看,则需要通用信息处理系统。

矛盾的解决方法之一就是要采用“中西文兼容技术”。如何实现“兼容”,“兼容”到何种程度?计算机体系结构、输入输出设备、系统软件、应用软件等又需怎样改造?是否要为中文信息处理设计新的软硬件系统?

我们主要是去解决所谓“计算机中文化”,还是“中文计算机化”?其关键何在?目前中文信息处理系统开发应用水平如何?今后发展方向是什么?

这些问题本书都作了一些探讨。本书的作者还提出了不少独特见解,如中

文信息处理系统五层结构模型,中文信息处理系统输入输出映射关系,通用输入输出模块结构思想,中文信息处理系统的 T-MC-C(终端—微型机—计算机)三级结构,中文信息处理码系列及与交换码之关系等等。

本书作者在近年繁重的科研任务下,投入了艰巨的劳动,经过多方深入的调查研究,形成了作者本人的设计思想方法,获得广大同行的注意。

作者在撰写学位论文时,就表现出有独立钻研和开创工作的能力。我们一开始就支持他将多年积累的计算机中文信息处理的资料和心得,编写成教材,在中国科学院研究生院为研究生讲课。以后这本书的内容又被许多高等院校选作为教材。现在他将此书整理出版,我们将这本书推荐给广大读者。相信它的出版,能引起从事计算机中文信息处理系统研制,开发和应用的专家和学者们的重视,从而对推动我国计算机和中文信息处理事业的发展,起到积极的作用。

中国科学院计算技术研究所

范新弼 教授 博士

一九八五年八月十日

电子工业部

陈力为 教授

一九八五年八月二十五日

JS335/17

Our ancestor created Chinese Characters and invented the technology of paper-making and the movable-type printing technics, which made it possible that we enjoy a glorious and brilliant culture today.

The invention of electronic computers had epoch-making significance for the development of science and culture of the world. Since then computer has undergone four generations in its development and is now advancing to a new stage —intelligent stage. One of the important features of the new trend is to adopt natural language or approximate natural language as means for man-machine interaction. In China this means using Chinese.

Electronic computer was invented in the west, therefore it is very well fitted to the Alphabetic letters of the western languages. Alphabetic letters are small in number and simple in structure. However, Chinese ideographs are large in number and quite complex in structure. Now we have the problem, how to use electronic computer to deal with the information in Chinese?

There exist two points of view to deal with the computer Chinese information processing. The first one is to modify the available computer system in order to make it fit the characteristics of Chinese information. This is the problem in terms of "computer processing of Chinese information", the other one is to reform the Chinese characters so as to adapt them for use in the current computer processing system. This is the problem of "the adaptation of Chinese information for computer processing". To solve these problems will require the joint efforts of not only computer scientists, linguists and philologists, but also mathematicians, electronics scientists, biologists, psychologists, historians, logicians and etc.

Form, sound and meaning are the three essential elements of Chinese ideographs. The characteristics of its pronunciation is one syllable, and Chinese characters are in the form of two-dimensional structure. In addition, Chinese characters are large in number, various in style and type. All these make it difficult to adapt Chinese ideographs to the current computer code system and hardware and software structure.

Besides, the computer's one direction letter processing does not fit the Chinese character's arrangement (horizontally from left to right and from top to bottom, while vertically from top to bottom and from right to left), the Mongolian's arrangement vertically from top to bottom and from left to right) and the Uygur's and Kazak's arrangements (horizontally from right to left and from top to bottom).

From the viewpoint of our philologists, a specific Chinese information processing system is in need. However, taking into consideration of the more broad view of world science, technic and culture development, a universal information processing system will be in great demand.

So, it seems to be that one of the best solution of the problem is to design a computer system with compatibility to both Chinese and Western languages. But how to realize the compatibility and to what degree? And how to modify the computer architecture, input and output device, system

software and application software? Is there a need to design a new software and hardware system for Chinese information processing?

In this book, after presenting the basic Knowledge of Chinese information processing, the author discusses all the prementioned problems and also explores the following problems: Shall we solve the problem by means of adapting the computer to Chinese characters or adapting the Chinese characters to the computer? What is the key in solving these problems? What is the status quo of the Chinese information processing? And what is the trend of its development? Moreover, the author also presents in this book some new ideas such as the five-layer structure model for the Chinese information processing system, the mapping relationship between input and output in the Chinese information processing system, the structure idea of general purpose module for input and output, the T-MC-C (Terminal-Microcomputer -Computer) three-stage structure of the Chinese information processing system, and the series of the Chinese information processing code and its relationship with the interchange code.

In recent years, the author has been engaged in heavy scientific research work. And only after making great effort and thorough investigation has he come to this new and integral view of Chinese information processing system, which has drawn great attention from his colleagues and co-workers in this field in China.

From the very beginning we supported him in compiling a book, in which he can summarize his experience and put in the Chinese information processing data he had accumulated for years of hard work, and he finished the book in 1983. Since then he has been using it as a textbook in the Graduate School, Academia Sinica. In recent years this book has been also taken as a textbook for students in computer science department of many universities.

At the time of the publication of this new revision, we sincerely recommend this book to our readers and strongly believe that its publication will draw the attention of experts and scholars in the field of research, production, development and application of the computer Chinese information processing. We also hope that it will play an active role in the development of computer and its Chinese information processing in China.

Professor, Ph. D Fan Xin-bi

Professor, Chen Li-wei

序 文

我々の祖先は漢字をつくり、紙と印刷活字を発明した。それ故我々は今日の輝かしい文化と文明を持っている。

電子計算機の発明は人類の科学文化生活に対し画期的な意義を持つ。電子計算機は四段階の発展を経て現在は新時代すなわち人工知能時代に向って進展している。新時代の計算機の知能表現の一つは、すなわちマンマシン・インターフェイスとして自然言語や自然に近い言語の採用であり、我が国ではそれは中文の採用である。

電子計算機は欧米で生れたので、そのハードウェア・ソフトウェアシステムはアルファベット表音文字が適用されている。アルファベット表音文字は文字数も少く、構造も簡単である。しかし中文なかでも漢字は数が多量で構造も複雑である。以何に計算機を利用して中文情報処理をすべきだろうか！

現在計算機で中文情報処理をするには二つの問題がある。一つは以何に現存の計算機を改造し、それを中文情報の特徴に適應させるか、すなわち“計算機の中文情報処理”の問題である。いま一つは以何に中文を改革すれば、それが計算機処理に適應するか、すなわち“中文情報の計算機処理”の問題である。これらの問題を徹底的に解決するには計算機の専門家に頼るばかりではいけないし、言語学専門家ばかりでもいけない。数学、物理、電子、生物、心理、歴史、編集等の各分野の専門家が必要である。

漢字の三要素は形、音、意味である。漢字の発音の特徴は単音節である。漢字の形は二次元パターンである。漢字の字数は多く、字体も多く、字号も多い。これらはみな現計算機のコード体系、ソフト、ハードと互いに構造が矛盾している。

計算機の文字順序処理は、中文情報の漢字は左から右上から下への横書き配列、上から下右から左の縦書き配列、モンゴル語の上から下左から右への縦書き配列、ウイグル語(ハザク文字)の右から左上から下への横書き配列等の特徴と相入れない。

言語学の角度からみると専門的な中文情報処理システムが必要である。世界の科学技術文化共通性からみると汎用情報処理システムが必要である。

矛盾したものの解決方法の一つとして中文欧米文互換技術を採用すべきである。以何に互換を実現するか、互換ほどの程度にするか？計算機システム構成や入出力装置やシステムソフトウェアやアプリケーションソフトウェア等をまたどのように改造する必要があるか？中文情報処理の為に新しくソフト・ハードシステムを設計する必要があるだろうか？

我々は主にいわゆる“計算機の中文化”か、それとも“中文の計算機化”か？そのキーボ

イントはどこにあるか？現在の中文情報処理システム開発の応用水準はどうか？今後の発展の方向は何か？と解決してきた。

これらの問題は本書がすべて検討している。本書の作者は独自の見解として、中文情報処理システムの五層構造モデル、中文情報処理システム入出力マッピング関係、汎用入出モジュール構造思考、中文情報処理システムのT-MC-C三レベルのアーキテクチャ、中文情報処理コードモードと交換コードの関係等を少からず提起している。

本書の作者は近年多忙な科学研究業務に精力的な作業をやり、深く多く調査研究し、作者本人の設計思想方法を形成して同分野の多勢の者に注目されている。

作者は学位論文を書いた時も独自の研究と創造力を表現していた。我々は当初より彼が多年蓄積してきた計算機中文情報処理の資料と経験を教材に編集させ、中国科学院研究生院で研究生の為に講義した。その後この本の内容はまた多くの大学研究院で教材に収められた。今般著者がこの本を整理出版するにあたり、我々はこの本を多くの読者に推薦するものである。この出版が計算機中文情報処理システムの研究、開発や応用の専門家、学者達に重視され、それが我が国の計算機と中文情報処理事業の発展を推進するために積極的な衝きを引き起こすことと信じている。

範新弼

中国科学院研究員

一九八五年八月十日

陳力為

電子工業部教授

一九八五年八月二十五日

Nos ancêtres ont inventé les caractères chinois, la papeterie et l'imprimerie typographique. Aussi possédons-nous notre brillante culture et civilisation d'aujourd'hui.

L'invention de l'ordinateur a pour la vie scientifique et culturelle de l'humanité une signification qui fait époque. L'ordinateur a connu quatre générations. Actuellement, la nouvelle étape d'ordinateur, celle d'intelligence, est en train d'apparaître. Une des caractéristiques d'intelligence important est le dialogue entre machine et homme sous le langage naturel ou langage approximativement naturel qui est le chinois en Chine.

L'ordinateur est né à l'Occident, son système du software et du hardware convient aux écritures phonétiques de l'Occidental. Dans ce langage, le nombre de lettres phonétiques est petit et la structure est beaucoup plus simple, par contre, le chinois idéographique possède de très nombreuses lettres et sa structure est bien compliquée. Alors comment pourrait-on traiter les informations en chinois avec un ordinateur?

Actuellement il existe deux problèmes pour cela: le premier est de modifier les ordinateurs existants pour les rendre convenables au traitement des informations en chinois, c'est-à-dire, le problème du "traitement des informations chinoises par l'ordinateur". Le deuxième est de réformer le chinois pour le rendre convenable au traitement des informations par l'ordinateur, c'est-à-dire, le problème de "l'adaptation pour le traitement par l'ordinateur des informations chinoises". Pour résoudre définitivement ces problèmes, on ne doit pas s'appuyer seulement sur les scientifiques d'ordinateur, ou les linguistes et philologistes, mais aussi sur d'autres spécialistes, tels que les mathématiciens, les physiciens, les électroniciens, les biologistes, les psychologues, les historiens, les médecins, etc.

La forme, le son et la signification constituent les trois éléments essentiels du chinois. La particularité du son du chinois est monosyllable. Les caractères chinois sont de graphiques de deux dimensions. Leur nombre, leur forme et leur grandeur typographique sont respectivement variés. Tout ce qui est décrit au-dessus est contradictoire au système existant de code de l'ordinateur et à la structure du software et du hardware.

Les caractères chinois, traités dans les informations, s'écrivent horizontalement de gauche à droite et de haut en bas, ou verticalement de droite à gauche et de haut en bas, et le mongol s'écrit verticalement de haut en bas et de gauche à droite; et l'ouïgour ou le kazakh s'écrivent horizontalement de droite à gauche et de haut en bas. Tout cela est contradictoire au traitement suivant l'ordre des caractères par l'ordinateur.

Du point de vue de linguistique et de philologie, on a besoin d'un système spécial pour le traitement des informations chinoises. Prenant en considération de la "coexistence" de science, de technique et de culture mondiale, il est nécessaire d'avoir un système universel du traitement des informations.

Une des méthodes de la solution de la contradiction est de mettre en application la "technique de

compatibilité du chinois et de la langue Occidentale". Comment se réalisera cette technique? A quel degré se réalisera-t-elle? Comment se transformeront l'architecture de l'ordinateur, les dispositifs d'entrée et de sortie, le software systématique et le software d'application? Faudra-t-il dresser le projet d'un nouveau système du software et du hardware pour le traitement des informations en chinois?

Devrons-nous résoudre pour l'essentiel le problème soi-disant de "rendre l'ordinateur utilisable pour le chinois" ou bien de "rendre le chinois convenable à l'ordinateur"? Quel est le noeud du problème? Comment se trouve le niveau de développement et d'application du système de traitement des informations en chinois et vers quelle direction se dirige-t-il?

Des études sur tous ces problèmes ont été faites dans ce livre. L'auteur présente bien des points de vue particuliers, par exemple: le modèle de structure de cinq niveaux du système de traitement des informations en chinois, l'idée de structure universelle de modèle pour les entrées et sorties, les relations biunivoques entre les entrées et sorties au système de traitement des informations en chinois, la structure de trois étapes T-MC-C (terminal-microordinateur-ordinateur) du système de traitement des informations chinois, le système de code du traitement des informations chinois et la relation entre ce système et les codes d'échange, etc.

Depuis ces dernières années, l'auteur, en faisant ses nombreuses recherches, a fait des travaux difficiles et de profondes investigations. C'est ainsi que sa méthode idéologique de conception a formée. Cela a donné une forte impression à ces collègues.

Quand il a rédigé sa thèse de grade, l'auteur montrait déjà sa capacité de faire indépendamment les recherches et son esprit d'initiative. Nous l'encourageait à rédiger un manuel aux moyens des matières sur le traitement des informations en chinois par l'ordinateur, qu'il avait accumulées pendant bien des années. L'auteur a donné des cours à l'Ecole des Aspirants chercheurs de l'Académie des Sciences de Chine. Le contenu de ce livre a été choisi par des écoles supérieures comme matière d'enseignement.

Maintenant, l'auteur a rédigé une nouvelle édition de livre et la rendu à la publication. Nous présentons ce livre au grand lecteur. Nous croyons que la publication de ce livre puisse attirer l'attention des spécialistes et des savants qui s'occupent de l'étude du développement et de l'application du système du traitement des informations en chinois par l'ordinateur; et nous sommes pleins d'espoir qu'elle prenne un rôle actif dans le développement du traitement des informations en chinois et de l'ordinateur de Chine.

Professeur, Dr. Fan Xien — bi

Professeur, Chen Li — wei

Предисловие

Наш предок создал китайские иероглифы, изобрел технику бумажного производства и технику печатанию. в следствии этого появилась блестящая китайская культура и цивилизация.

Изобретение электронных вычислительных машин (ЭВМ) открыло новую эру в истории развития науки и культуры. ЭВМ прошла целые четыре поколения и вошла в новое поколение, при котором она представляет собой как умственное устройство. Одной характерной чертой такого устройства является интерфейс "человека-машины" через естественный язык, т. е. китайский язык в нашей стране.

ЭВМ впервые появилась на западе. Ее аппаратные и программные средства соответствуют западным алфавитам, которые имеют немногие буквы и простую структуру. Количество китайских иероглифов гораздо больше и их структуры гораздо сложнее. Естественно возникает такой вопрос, каким образом использовать распространенные эвм для обработки информации на китайском языке.

При применении ЭВМ для обработки информации на китайском языке имеются два вопроса. Первый из них заключается том, каким путём надо преобразовать ЭВМ, чтобы они соответствовали чертам информации на китайском языке, т. е. "вопрос об обработке информации на китайском языке подходящимися машинами". Второй вопрос такой, как реформировать китайские иероглифы, чтобы они были удобными для обработки машинами, т. е. "вопрос об обработке информации на китайском языке существующими машинами".

Для окончательного решения таких вопросов требуются не только специалисты по ЭВМ и лингвисты, но и математики, физики, электроники, биологи, психологи, историки, логики и другие.

Китайские иероглифы обладают тремя факторами: формой, произношением и значением. Они имеют двухмерную форму и различные шрифты. Их особенностью произношения является монослог. Всё это находится во противоречии со системой кодов, аппаратными и программными средствами существующих ЭВМ.

ЭВМ в общем последовательно обрабатывают одномерные знаки. Они не смогут принимать информации некоторых наций. Например, информация на китайском языке читается порядком слева направо и сверху вниз; монгольская информация пишется сверху вниз и слева направо; казачая информация имеет порядок справа налево и сверху вниз.

С точки зрения иероглифов требуется специальная ЭВМ для обработки информации на китайском языке, а с точки зрения использования совместного богатства науки, техники и культуры для этой цели лучше применять распространенные машины.

Одним способом решения таких противоположных вопросов является применение техники совместимости между информацией на западных языках и информацией на китайском языке. Как и до какой степени можно осуществить такую совместимость? Каким образом нужно преобразовать конструкцию машины, её устройства ввода-вывода и программное средство? Нужно ли

проектировать новую ЭВМ и новое программное средство? Что главное, машина должна соответствовать информации на китайском языке или наоборот? Что представляет собой существенное? На каком уровне находится настоящая работа по обработке информации на китайском языке? и по какому направлению она должна развиваться?

Все эти вопросы будут рассмотрены в этой книге. Автор предлагает многие уникальные мнения, как пятиступенчатую конструктивную модель системы обработки информации на китайском языке. Отражение между вводом и выводом, идею модульной конструкции устройств ввода и вывода. Т-МС-С трехступенчатую систему, формат кодовой системы обработки, отношение между такой системой и кодовой системой обмена и т. д.

При тяжелой исследовательской работе за последние годы автор тратил большой труд, формировал свою проектную идею и вызвал большое внимание от специалистов и коллег.

При защите диссертации автор выразил его изумительную и творческую способность. Мы оказали ему поддержку, чтобы он систематизировал накопленные материалы и написал лекции для академии аспирантов. В настоящее время его лекции, как учебник, используются многими вузами. На этом основании автор написал эту книгу.

Мы рекомендуем её нашим читателям и верим, что на неё большое внимание будут обращать специалисты и ученые, работающие в области обработки информации на китайском языке. Эта книга будет играть несомненную роль для развития техники по обработке информации на китайском языке электронными вычислительными машинами.

Профессор, фан Синби

Профессор, Чжен Ливей

VORWORT

Die Kreation der chinesischen Schrift, die Erfindung des papiers und der Druckkunst mit beweglichen Lettern durch unsere Vorfahren haben zu unserer heutigen hochentwickelten und glorreichen Kultur und Zivilisation geführt.

Die Erfindung der elektronischen Datenverarbeitungsanlagen hatte auf das wissenschaftliche und kulturelle Leben der Menschen eine epochale Auswirkung. Der Computer erlebte vier Entwicklungsstufen und schreitet weiter zur neuen Generation der Intelligenz. Eine der Eigenschaften der Computerintelligenz besteht im Dialog zwischen Menschen und Computer, der in natürlichen oder naturähnlichen Sprachen geführt wird. In China wird er auf Chinesisch abgewickelt. Der Geburtsort des Computers liegt in Westen, das Hard- und Softwaresystem ist auf die Verwendung westlicher Schrift festgelegt. Ihr Alphabet enthält nur wenige Buchstaben, ihr Schriftbild ist relativ einfach. Die chinesische Schrift besteht aus einer großen Anzahl von Schriftzeichen mit kompliziertem Schriftbild. Wie kann sie mittels Computer verarbeitet werden?

Gegenwärtig herrschen dazu zwei Ansätze: bei dem einen geht es um die Anpassung des Computers, wodurch es den Charakteristiken der chinesischen Sprache gerecht wird. Man versteht darunter die Umstellung des Computers auf die Verarbeitung chinesischer Information, bei dem anderen geht es um die Modifikation der chinesischen Schrift, um sie vom Computer einfach zu verarbeiten. Man nennt dies die Umstellung der chinesischen Sprache für die Datenverarbeitung. Zur Lösung des Problems braucht man nicht nur Computer- und Sprachwissenschaftler, sondern muß sich auf die Mithilfe von Mathematikern, Physikern, Elektronikspezialisten, Psychologen, Historikern, Logikern und anderen Wissenschaftlern stützen.

Die drei Hauptelemente der chinesischen Schrift sind Form, Laut und Bedeutung. Der Laut ist einsilbig; die Schrift stellt sich in zweidimensionalen Bildern dar, sie enthält eine große Anzahl von Schriftzeichen, viele Stile und Typen. All das läßt sich schwer mit den vorhandenen Codesystemen, Hard- und Softwarestruktur des jetzigen Computers vereinbaren.

Die Anordnung der Zeichen in eine Richtung durch den Computer eignet sich kaum für die Darstellung chinesischer Schriftzeichen, die horizontal von links nach rechts und von oben nach unten, vertikal von oben nach unten und von links nach rechts verlaufen. Man muß noch die Darstellung von mongolischen Schriftzeichen, die vertikal von oben nach unten und von links nach rechts verläuft, und der Uiguren und Kasaken, die horizontal von rechts nach links und von oben nach unten verläuft, berücksichtigen.

Aus der Sicht der Sprachwissenschaft benötigt man ein spezifisches System für die Verarbeitung chinesischer Information, jedoch die weltweite gemeinsame Entwicklung der Wissenschaft,

Technologie und Kultur erfordert ein universelles Informationsverarbeitungssystem.

Eine Lösung liegt in der Kompatibilität für die westliche und die chinesische Sprache. Hier gehtes um die Fragen: Wie kann man die Kompatibilität realisieren, bis zu welchem Grad ist sie möglich, wie sollen die Architektur des Computersystems, die Ein- und Ausgabegeräte, die System- und Applikationssoftware modifiziert werden, wird für die Verarbeitung chinesischer Information neues hard- und Softwaredesign benötigt?

Sollen wir den Schwerpunkt auf die Sinisierung des Computers oder auf die Computerisierung der chinesischen Schrift legen? Wo liegt der Schlüssel zur Lösung des Problems, wie ist der Stand der Verarbeitung chinesischer Information gediehen, wohin geht der Trend der Entwicklung?

Alle diese Fragen werden in dem Buch zur Sprache kommen. Der Autor präsentiert darin viele eigene Anschauungen; wie z. B. das Fünfschichtenmodell der VerarInformation, die Zuordnungsrelation der Einund Ausgabe, die Idee der Konstruktion von Universalmodulen für die Einund Ausgabe, die Dreistufenstrukture (Terminal-Mikrocomputer-Computer) des Verarbeitungssystems für die chinesische Sprache, das Codesystem für die Verarbeitung der chinesischen Schrift und seine Beziehung zum Umwandlungscode.

Der Autor hatte in den letzten Jahren schwierige Forschungsaufgaben zu bewältigen. Durch seine Engagement für die Probleme der Verarbeitung der chinesischen Sprache konnte er eigene Gedanken entwickeln und die Aufmerksamkeit der Fachkollegen auf sich ziehen.

Zur Zeit seiner Diplomarbeit zeigte der Autor bereits die Fähigkeit selbstständiger Forschung undeigenen Schaffens. Wir unterstützten ihn vom Anfang an, seine mehrjährige Erfahrungen in der Verarbeitung chinesischer Information in einem Lehrbuch niederzuschreiben. Später wurde es im postgraduierten Institut der Akademie der Wissenschaften und in einigen Hochschulen verwendet. Aus diesem Lehrbuch entstand dann dieses Werk, das wir einem breiten Leserkreis empfehlen. Wir sind überzeugt, daß die Herausgabe dieses Buches das Augenmerk der Experten in der Forschung, Herstellung und Verwendung von Computersystemen für die Verarbeitung chinesischer Information auf sich lenken wird und eine positive Auswirkung auf die Entwicklung der Computerisierten Verarbeitung der chinesischen Sprache haben wird.

Prof. Fan Xinbi

Prof. Chen Liwei

前言

八十年代初期,随着计算机的推广应用与普及,对计算机能处理中文信息的要求越来越迫切,越来越高。在我们的科研工作上,有机会研究计算机中文信息处理问题。中文信息处理特别是汉字信息处理,不仅仅是输入输出问题,还有计算机体系结构、系统软件、数据库技术、网络通信、应用软件以及模式识别等一系列问题,需要我们炎黄子孙来解决,这是中国要实现计算机化的关键问题。

一九八二年十一月,在郑州中国电子计算机学会第六届年会期间,中国科学院研究生院罗晓沛副教授找我谈话,希望能为研究生院开设“计算机中文信息处理”课程。从内心上说,感到这是一项光荣而艰巨的任务。由于我们从未讲过课,同时国内还没有一本有关计算机中文信息处理的图书、教材或讲义,更不知这门课应包括哪些内容,害怕不能胜任,压力很大。

我们的导师、计算机界老前辈、中国科学院计算技术研究所研究员范新弼先生热情鼓励、积极支持我们开课。在北京信息控制研究所副所长王正中、于景元高级工程师支持下,才下决心应罗副教授邀请开课。

当时正值LS—83中文图形信息处理系统总体方案设计阶段。科研任务很重,只能抽业余时间备课,写教材。一九八三年五月完成了《中文信息处理(讲义)》初稿(共六册)的编写。一九八三年九月,正式在中国科学院研究生院开设“计算机中文信息处理”课程,受到了各界人士的欢迎,收到了良好效果,讲义也供不应求,这对我们的工作鼓舞很大。

一九八四年四月,研究生院决定将《计算机中文信息处理》正式列为硕士学位课程。我们在LS—83中文信息系统紧张调试的情况下,抽空对初稿进行了较大修改,形成《计算机中文信息处理》讲义第二稿(共四册)。空军第三研究所副总工程师张佑鸿以及刘湘江、刘文平同志对第二稿审阅、校对、发行做了大量工作。同年十月起陆续出版了共四册的第二稿讲义。向全国,除台湾省和西藏自治区外的二十八个省市自治区发行。先后被全国十多所高等院校选作计算机专业研究生和本科生教材。也被中国中文信息研究会选作培训课本,先后举办了若干期培训班。

一九八四年五月,宇航出版社廖寿琪高级工程师鼓励我们在第二稿基础上,再充实提高,并且提出具体改意见,以便使本书正式出版、发行、以满足国内外科学技术发展的需要。他还在整个组稿过程中,对稿件进行了认真、细致地审核,使我们的工作能顺利地进行。

在一九八四年九月至一九八五年元月研究生院第二次授课之后,听取了各方面专家、学者的宝贵意见,并结合科研工作成果,着手第三稿整理。

一九八五年春,国务院电子振兴领导小组顾问,中国中文信息研究会副理事长陈力为教授和中国中文信息研究会常务理事石云程副总工程师热情赞扬《计算机中文信息处理》讲义的第二稿,并对第三稿提出了许多指导性意见。中文信息研究会姜德存、龚滨良、袁琦等工程师也都给予多方支持。在此一并致以谢意。

一九八五年九月至一九八六年元月,第三次在研究生院讲授《计算机中文信息处理》,并对第三稿部分内容进行试讲。

在此要特别指出的,在我们从事计算机中文信息处理系统研制,教学以及本书的三年写作过程中,自始至终都得到了国务委员、国家科委主任宋健教授的关怀和鼓励,在此深表敬意。

衷心感谢为本书撰写书名的中国书法协会主席、北京师范大学启功教授,他不辞高龄,挤出休息时间提汉字书名。还要感谢中国中文信息学会少数民族语言文字专业委员会副主任嘎日迪、中央民族学院副院长胡坦教授、新疆大学副教授吾守尔、新疆大学副校长沙比汗副教授、延边大学讲师金淑子、中央民族学院教师岭福祥等同志,他们在百忙工作中分别为本书提蒙、藏、维、哈、朝、彝文的书名。也要感谢中国语言文字工作委员会研究员高景成同志,他用中国最古老的甲骨文为本书撰写书名。

感谢为本书写序的二位计算机界老前辈范新弼先生和陈力为先生。

在工作和写作的过程中,陈耀星、张建平、郝卫、郑爱华、吴泓、赵伯毅、张原、刘志刚等同志给予各种帮助,还有许多计算机界老前辈、专家和学者也都提出了许多宝贵意见,在此表示感谢。

为了与国外读者加强中文信息处理技术的交流,特将本书的序译成英、日、法、俄、德文,参加译校工作的有韩文盛、张爱英、关烈贤、张广仁、赵厚君、鲁小文、马淑雅等同志,奥籍华人王剑生博士在百忙的工作中为本书序译成德文,并将译稿通过传真和邮寄两种方式从维也纳寄到北京,在此对他们的辛勤劳动表示深切的感谢。

在本书的编辑、出版过程中,宇航出版社、铁道出版社印刷厂和国防工业出版社经过友好协商,决定加强横向联系,将本书作为合作项目。因为这套书(包括上、下册)从排版方面说非常复杂难排,既有图、表各数百幅,又有五种外国文字,也有程序、数学公式、上下角标、上或下带符号的字符,还有汉字拆成不成字的偏旁、部首、字元,甚至在古今中外汉字库中也未曾见到过的汉字等等。须要造字的数量大,采用国产华光 III 型照排系统的科技版进行试排,也不能完全满足需要,特别是一些图表连续排的情况下,要采取人工干预的方法。江啸高级工程师、沈珠工程师、刘润松助理工程师、李泽民等同志在技术上采取了一些措施,而且组织人力日以继夜地工作,没有他们的辛勤劳动,本书不可能在这样短的时间内出版,在此对他们的工作深表敬意。

全书共分十四章,每章主要内容如下:

第一章,探讨计算机中文信息处理可涉及的范围,建立了中文信息处理系统五层结构模型和输入输出映射关系。

第二章,简述现代汉语基础知识。

第三章,介绍了形码、音码、形音码和音形码等汉字输入编码原理并列举了编码例子。讨论汉字键盘输入编码评测方法有关事宜。

第四章,论述了汉字字形库结构、汉字点阵字型设计有关问题。

第五章,在复习各种检索技术基础上,详细阐明了通用汉字输入模块设计方法。