



[12] 发明专利说明书

[21] ZL 专利号 96103701.6

[43] 授权公告日 2003 年 3 月 5 日

[11] 授权公告号 CN 1102779C

[22] 申请日 1996.3.21 [21] 申请号 96103701.6

[30] 优先权

[32] 1995. 3.24 [33] JP [31] 66117/1995

[71] 专利权人 松下电器产业株式会社

地址 日本大阪府门真市

[72] 发明人 郭俊桔

[56] 参考文献

CN1038364A 1989.12.27 G06F15/38

CN1045878A 1990.10.03 G06F3/16

JP4238397A 1992.08.26 G10L3/00,G10L5/00

审查员 盖 浩

[74] 专利代理机构 中国专利代理(香港)有限公司

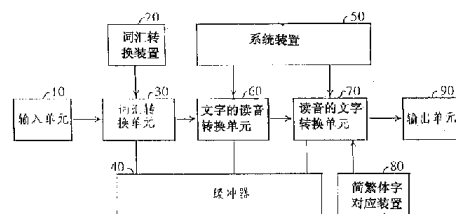
代理人 王 勇 叶恺东

权利要求书 1 页 说明书 10 页 附图 12 页

[54] 发明名称 中文简繁体字文件转换装置

[57] 摘要

提供了一种正确转换中文简体文件和繁体文件的装置。除汉字的文字及读音信息以外，还参照词汇转换装置。具体地说，词汇转换单元将原始文件及其语言标识存入缓冲器并参照语言标识及词汇转换装置改写汉字与单词。文字的读音转换单元参照语言标识和系统装置将原始文件的文字转换成读音符号。读音的文字转换单元参照语言标识、系统装置、简繁体对应装置通过从读音到文字的转换算法将原始文件的读音转换为目标文件的文字并且输出。



1. 一种中文简繁体字文件转换装置，这种装置采用预定的语言标识来区别中文简体字和繁体字，该装置在将用中文简体字或繁体字记录的原始文件转换为用另一种文字记录的目标文件，所说的中文简繁体字文件转换装置配备有：

存储简体字词汇及与其对应的繁体字词汇的词汇转换装置；存储读音符号和与其对应的简体及繁体汉字或单词的系统装置；

存储简体字以及与其对应的繁体字的简繁体字对应装置；

10 以输入的语言标识来鉴定原始文件是简体字或繁体字，再以输入的原始文件词汇检索上述的词汇转换装置，找出适当的对应字，改写原始文件词汇的词汇转换单元；

参照上述系统装置将上述词汇转换单元所产生的文件中的文字转换为读音符号的文字的读音转换单元；

15 参照上述系统装置和简繁体字对应装置，通过从读音符号到文字的既定转换算法将上述读音符号转换为其它字体的目标文件文字的读音的文字转换单元，

其特征在于：简繁体字之间的转换兼用了文字及读音的信息。

2. 如权利要求 1 所述的中文简繁体字文件转换装置，其特征在于：前述的系统装置具有无论中文的汉字和单词是简体字还是繁体字，都将非变音字存储在非变音字区的非变音字区装置以及将变音字存储到变音字区的变音字区装置；前述读音的文字转换单元具有采用最长一致法作为转换算法的最长一致法反映转换装置。

25 3. 如权利要求 1 或 2 所述的中文简繁体字文件转换装置，其特征在于：作为转换算法，前述读音的文字转换单元具有将使用频度高的文字和单词优先进行转换的使用频度反映转换装置。

4. 如权利要求 3 所述的中文简繁体字文字转换装置，其特征在于：前述按使用频度的转换装置具有切换与原始文件内容相应的上述转换算法或切换转换时使用的频度表的按文件内容的使用频度转换控制装置。

中文简繁体字文件转换装置

5 技术领域

本发明涉及中文简体字文件和繁体字文件的转换装置。

背景技术

近几年来，中国大陆和台湾之间的交流频繁，两者间的文件来往也随之增加，但是，由于长达 40 年间没有交往，两地不仅使用的汉字字体不同，就是单词和词汇（日常使用的单个或多个单词的使用方法）等也有很大不同，因此，难以相互理解对方使用的文件。例如，台湾将 laser printer 叫作“雷射印表机”，而大陆则称之为“激光打印机”。因此，大陆使用的简体字文件和台湾使用的繁体字文件之间相互转换的需要大量增加，在简体字文件和繁体字文件间的相互转换中，需要解决以下的技术和语言方面的困难。

(1)简体字的常用字数大约是 8000 个，而常用的繁体字比常用简体字多。例如，计算机领域的常用繁体字就有 13,053 个。所以，会有若干个繁体字对应于一个简体字的情况，如简体字的“后”与繁体字的“后”和“後”相对应。因此，在从简体字文件转换至繁体字文件的情况下，需要选择适当的繁体字。

(2)由于社会、文化的差异，单词及其用法也不相同。例如，用于表现日常所说质量程度的“水准（技术水平等）”，中国大陆称为“水平”，而台湾则称为“水准”。

所以，研制开发了用简体字写的文件和用繁体字写的文件的转换装置。

先有中文简繁体字文件转换装置有诸如中华民国 1994 年第 7 次计算机语言学研究会论文集第 187 至 201 页的“A Text Conversion System Between Simplified and Complex Chinese Characters Based on OCR Approaches”所述的装置。该论文说明了从简体字文件到繁体字文件转换装置的实例。图 1 显示了这种装置的结构。在该图中，标号 100 是可以输入以简体字文件或繁体字文件作为原始文件的原始文件输入单元。标号 300

是存储各个文字使用频度的文字使用频度表。标号 350 是存储每个文字特征值的特征数据库。标号 200 是从图象数据中抽取出文字的文字切取单元。标号 210 是计算并抽取从图象数据中所选出的文字的特征值的特征抽出单元。标号 220 是参照文字使用频度表 300 及特征数据库 350, 进行文字对比的对比单元。标号 500 是存储单词的单词转换装置。标号 510 是存储一般文字信息的基本装置。标号 520 是存储简体字及繁体字文字代码的代码对应表。标号 530 是存储事先根据统计等求出的相邻文字之间连接频度的 BIGRAM 表。标号 400 是将候补汉字或单词转换为文字网络(Word lattice: 指正在处理的各候补的汉字、单词等形成的网络结构)的中文语言装置, 它具有单词等转换装置和文字修正装置。标号 420 是寻找最佳转换路径的汉字和单词转换单元。标号 410 是通过人工等方法修正误识别文字的文字修正单元。标号 600 是输出进行转换后所得目标文件的输出单元。

下面以图 2(a)所示的简体字文件为例, 说明在以上中文简繁体字文件转换装置中从简体字转换到繁体字的转换顺序。

一旦由原始文件输入单元 100 输入图 2(a)所示的简体字文件, 就通过 OCR (光学字符读取装置) 将作为图形的图象读入 OCR 装置。并且, 在通过文字切取单元 200 抽出各文字的图象之后, 由特征抽出单元 210 计算出各文字的特征值。利用对比单元 220, 参照文字使用频度表 300 以及特征数据库 350 检出候补的简体字。然后进行文字转换及后处理装置的操作。首先, 在中文语言装置 400 中参照代码对应表 520 取出如图 2(b)所示的各个相应的候补字。将已经取出的各候补的目标汉字组合作为检索键, 检索单词转换装置 500 及基本装置 510, 取出图 2(c)所示的候补单词。按照图 2(d)所示的文字网络组织已经取出的候补单词。由汉字和单词转换单元 420 参照 BIGRAM 表 530 (根据语言资料库(CORPUS)相邻的两个文字或单词的使用频度), 根据 Statistical bigram Markov Language Model(采用统计式 BIGRAM 的马尔可夫语言模型)从文字网络中取出图 2(e)所示的最佳的、也即, 取出可能性最高的转换路径后, 由输出单元 600 输出。

下面简要说明获得最佳转换路径的方法。

利用 BIGRAM 的统计数据($P(C_i | C_{i-1})$ 和($P(C_i | S_i)$), 找出能使下述函数值为最大的路径。

$$\text{Max} - P(C | S) = P(C_1, C_2 \dots C_n | S_1, S_2 \dots S_n) \rightarrow \text{Sum}. P(C_i | C_{i-1})(C_i | S_i) (i=1, \dots, N)$$

SS:候补汉字集,例如将第一个读取的候补集设为 S_1 。

C:候补集 S 的元素集,例如, C_1 表示 S_1 的第一个候补汉字。

$P(C_i | C_{i-1})$:表示一个字是 C_i ,而且这个字之前的字为 C_{i-1} 时在 BIGRAM 中出现的概率。

5 用相同的方法将上述公式应用于下列的公式。

→ Sum. $P(W_i | W_{i-1})P(C_i | S_i)(i=1, \dots, N)$

$SP(W_i | W_{i-1})$:表示一个单词是 W_i ,而且这个单词的前一个字为 W_{i-1} 时在 BIGRAM 中出现的概率。

但是,以上的内容已在 E. J. Yannakoudakis 和 P. J. Hutton 的载于 Pattern
10 Recognition 杂志 1990 年第 23 卷第 5 期 509 至 528 页的文章“n-Grams and
their implement to natural language understanding”的第 6 节中有所说明,并
且马尔可夫过程的算法如现代数学社 1981 年出版的《入门 | OR 讲座》第
96 页所说明的那样,都是众所周知的技术。所以省略说明。

图 3 表示以上顺序的流程。

15 上面已提到的先有技术存在以下问题:

(1)例如简体字的“于”经常被误识别为“干”,因而由 OCR 误识别的
文字难以通过后处理进行修正。

(2)由于只是利用文字信息,因而不能对中文的变音字进行有效处
理,所以,转换的正确性有一定的局限。这里的“变音字”是指有多种读
20 音的汉字。如果举出具体例子,那么就以简体字文件的“他干干净净的工作”
为例,“干”有“gan1”和“gan4”的读音,因此,由于繁体字的“乾”
和“干”(“gan1”)及“干”(“gan4”)相对应,因而,在先有技术中,“干
干”为(gan1)(gan1)所以,一般转换为“他乾乾净的工作”,而得不到
正确的“他干乾净的工作”。

25 (3)由于著作方式及写文章内容的不同,不易得到不同大规模并
较为平衡的中文语言资料库,所以,在抽取 BIGRAM 时需要许多工时和
大量经费。还有,如果中文语言资料库不均衡,抽出的 BIGRAM 表就会
影响单词转换的正确性。这里的语言资料库(CORPUS)是指与相邻的文字、
单词出现概率有关的数据库。

30 所以,需要有一种廉价且正确的中文简体字和繁体字文件相互转换的
装置。

发明内容

本发明的目的就是要解决这样的课题。

为了实现上述目的，本发明的中文简繁体字文件转换装置是用预定的语言标识来区别中文简体字和繁体字，例如前者使用“00”，后者用“01”来表示，本装置是一种利用文字及读音信息将用简体字或繁体字记录的原始文字转换为用另一种文字记录的目标文件的中文简体字文件转换装置，该装置特征在于，它配备有：词汇转换装置，此装置分组（包括实质的组）存储简体字词汇（包括单一或多个汉字、成语、语句等）以及与其对应的繁体字词汇；系统装置，此装置存储有读音符号（包括注音符号、拼音符号以及多种符号）以及与其对应的简体和繁体汉字（无论是单个汉字或多个汉字）或单词（包括词组、惯用句等）；简繁体字对应装置，此装置成组存储简体字及与其对应的繁体字；词汇转换单元，此单元用简体字或繁体字所记录的原始文件中的词汇去检索上述词汇转换装置，找出适当的对应词来改写原始文件的词汇；文字的读音转换单元，此单元参照系统装置将上述词汇转换单元所产生的文件中的文字转换为读音符号（包括各种字音符号）；读音的文字转换单元，此单元参照系统装置和简繁体字对应装置，根据从读音符号到文字的既定转换算法将上述读音符号转换为另一种字体的目标文件文字。

又者，上述本发明的中文简繁体文件转换装置，其中：前述系统装置带有将中文简、繁体汉字、单词的非变音字存储在非变音字区内的非变音字区装置以及将变音字存储在变音字区域内的变音字区装置（包括在同一存储器内的字用两者的标识加以区别）；前述读音的文字转换单元具有按最长一致法的转换方法，这种转换方法采用了最长一致法（同时也采用了所谓的形态要素解析法等其它方法），而最长一致法作为一种转换算法首先使构成的音节（汉字）数优先，其次使业已存在的音节优先。

再者，上述本发明之中文简繁体字文件转换装置，其中：上述读音的文字转换单元带有按使用频度的转换装置（包括其它装置），上述装置作为一种转换算法优先转换使用频度高的文字、单词。

又再者，上述本发明的中文简繁体字文件转换装置，其中：上述使用频度的转换装置带有按文件内容的使用频度切换的装置，此装置能够诸如根据领域的改变而切换与原始文件的技术性文件、文字性文件相应的转换算法或切换转换算法在转换时所用的频度表。

根据上述结构，本发明的中文简繁体字文件转换装置中，使用者通过

OCR、磁盘驱动器等输入原始文件及表示是简体或繁体的语言标识（包括实质相同的指示及输入）后，词汇转换单元用简（繁）体字原始文件的词汇检索事先存储了简体字词汇及与其对应的繁体字词汇的词汇转换装置，找出适当的对应字改变原始文件的词汇。系统装置预先存储了读音符号和与其对应的简体及繁体汉字或单词。文字的读音转换单元参照上述系统装置将原始文件的文字转换为读音记号。简繁体字对应装置预先成组地存储了简体字与其对应的繁体字。读音的文字转换单元参照系统装置和简繁体字对应装置通过从读音符号到文字的既定转换算法将上述读音符号转换为繁（简）体字的目标文件的文字。

10 又者，在上述本发明的中文简繁体字文件转换装置中，上述中文简繁体字文件转换装置的简体字和繁体字是用语言标识来加以区别的。无论中文的汉字、单词是简体字还是繁体字，系统装置都将非变音字和变音字分别划分为非变音字区和变音字区并加以存储（包括按文字代码顺序排列，采用其它标识的划分）。读音的文字转换单元采用作为转换算法的最长一致法进行转换。

15 再者，在上述本发明的中文简繁体字文件转换装置中，上述的中文简繁体字文件转换装置的读音的文字转换单元优先转换使用频度高的汉字和单词。

20 又再者，在上述本发明的中文简繁体字文件转换装置中，上述按使用频度的转换装置中的按文件内容的使用频度切换控制装置可以根据原始文件内容切换上述转换算法或用于转换的频度表。

下面根据实施方式说明本发明。

附图说明

- 25 图 1 显示了先有中文简繁体字文件转换装置结构的一个实例。
图 2 显示了先有中文简繁体字文件转换装置的处理流程。
图 3 显示先有中文简繁体字文件转换装置的一个处理范例。
图 4 是本发明之中文简繁体字文件转换装置的实施方式结构图。
图 5 是上述实施方式中的词汇转换装置的数据结构的概念图。
在本图中，(a)在从简体字文件转换成繁体字文件时使用。
30 (b)在从繁体字文件转换成简体字文件时使用。
图 6 是主要显示了上述实施方式中词汇转换单元操作的流程图。
图 7 是上述实施方式中系统装置数据结构的概念图。

图 8 是主要显示上述实施方式中文字的读音转换单元操作的流程图。

图 9 概念性地示出了上述实施方式中简繁体字对应装置的数据结构。

图 10 是主要显示上述实施方式中读音的文字转换单元操作的流程图。

5 图 11 是在上述实施方式从简体字文件转换为繁体字文件的具体实例中, 根据读音的文字转换单元图 10 的(c2)步骤(step)处理结果所取出的汉字和单词。

图 12 显示了根据图 10 的(c3)步骤处理结果所取出的与各汉字读音相对应的文字。

10 图 13 表示根据图 10 的 (c4)步骤处理结果所取出的与各汉字的读音相对应的文字。

具体实施方式

图 4 是与本发明有关的中文简繁体字文件转换装置的一个实施方式的构成图。

15 在该图中, 标号 10 是由与外部相连的通信线路 OCR, 它是通过磁盘驱动器等输入原始文件及语言标识的输入单元。标号 20 是存储图 5 示出的那种中国大陆和台湾之间词汇差异对应表的词汇转换装置。标号 30 是参照语言标识及词汇转换装置, 将与检索字一致的原始文件词汇改写为对应字的词汇转换单元。图 6 表示词汇转换单元 30 的处理流程。该图在后面作详细说明。标号 50 是存储各读音符号及与其对应的简体汉字和单词及繁体汉字和单词(包括变音字)的系统装置。系统装置 50 像图 7 那样分简体和繁体存储着汉字和单词。标号 60 是参照语言标识及系统装置 50 将原始文件的文字转换为读音符号的文字的读音转换单元。图 8 表示转换单元 60 的处理流程。该图也在后面详细说明。标号 80 是存储简体字和繁体字之间对应关系的简繁体字对应装置。标号 70 是参照简繁体字对应装置 80 将原始文件的读音符号转换为目标文件的读音的文字转换单元。图 10 表示读音的文字转换单元 70 的处理流程。该图将在后面详细说明。标号 90 是根据转换处理结果输出目标文件的输出单元。标号 40 是临时存储处理结果的缓冲器。

30 此外, 还有最后由翻译者进行校对的文字修正单元, 这些都是众所周知的技术, 所以省略图示和说明。

下面参照流程图说明本实施方式的操作。

首先说明图 6 所示的词汇转换单元 30 的操作。

(a1)在由输入单元 10 输入语言标识及原始文件的同时，将它们记录到缓冲器 40 中。

(a2)参照缓冲器 40 所记录的语言标识，按顺序从词汇转换装置 20 中取出检索字以及与其对应的置换字。

(a3)判断是否已全部取出了检索字。

在仍有检索字的情况下，进入 (a4)处理，将检索字作为检索键与原始文字的文字匹配后，返回到(a2)。

在没有检索字的情况下，结束词汇转换单元 30 的处理，进入文字的读音转换单元 60 的处理。

下面说明图 8 所示的文字的读音转换单元 60 的操作。

(b1)输入已由词汇转换单元 30 作了转换的原始文件的文字。

(b2)在根据区分符（例如文章中的逗号，句点）按每个音节切取出原始文件的文字后，将它们记录到缓冲器 40。

(b3)分别取出已经记录在缓冲器 40 中的各个音节，参照系统装置 50 将非变音字（不含混的字优先）的文字转换为读音符号后并记录到缓冲器 40。

(b4)对已由缓冲器 40 记录的有变音字的文字参照系统装置 50 的变音字区，将变音字转换为适当的读音符号。

(b5)参照缓冲器 40 的原始文件的文字，按照中文语法修正缓冲 40 的各个文字的读音符号。例如，“妈”的读音的“mal”，而“妈”的第 2 个“妈”的声调不按 1 声（在 4 声中最高）读，而是应按轻声（由于音节连续而失去了原有声调，弱且轻轻地发音）读为『ma0』，所以要修正第 2 个『妈妈』的读音符号。

关于文字的读音转换，由于是在诸如特开平第 4-238397 号于 1992 年 8 月 26 日所公开的众所周知的技术，所以省略了说明。

以上文字的读音转换单元 60 处理结束，就进入图 10 所示的的读音的文字转换单元 70 的处理。下面说明其内容。

(c1)输入由文字的读音转换单元 60 所得到的原始文件读音符号。

(c2)参照系统装置 50 切取出读音符号可能成为音节的全部音节。并将与各音节对应的候补汉字和单词记录到缓冲器 40。

(c3)参照缓冲器 40 所记录的原始文件的文字，将各文字作为检索

键，从简繁体字对应装置 80 中取出与各文字对应的文字，并存储到缓冲器 40。

(c4)从缓冲器 40 中取出可能成为音节的各音节候补汉字和单词，参照与缓冲器 40 中各读音符号相对应的文字，删除不适当的候补即删除可能性低的汉字和单词。

(c5)利用最长一致法，从候补的汉字和单词中选择适当的文字，据此，结束读音的文字转换单元 70 的操作。

下面以从中文简体字文件的原始文件，转换为作为目标文件的繁体字文件为例，较具体地说明本实施方式的操作。

10 若由输入单元 10 输入“他尚未使用软件”的原始文件及称为“简体”的语言标识，则词汇转换单元 30 要参照图 5(a)所示的词汇转换装置 20。由于有称为“软件”的检索字，所以有将原始文件的“软件”置换为对应字的“软体”，将原始文件改写为“他尚未使用软体”以后，将已经置换的原始文件及语言标识记录到缓冲器 40。然后，文字的读音转换单元 60
15 参照缓冲器 40 所记录的语言标识及图 7 所示的系统装置 50，将已改写的文字转换为下面所述的读音符号“ta1 shang4 uei4 sh3 yueng4 ruan3 ti3”。然后，进入图 10 所示的读音的文字转换单元 70 进行处理。若在(c1)中输入以上的读音符号，那么就在(c2)中切取音节，通过转换，取出一个有可能成为汉字、单词、词汇的音节。已经取出的音节以及与其对应的候补汉字和单词如图 11 所示。

在图 10 的 (c3) 中，参照图 7 所示的简繁体字对应装置 80 及缓冲器 40 中的原始文件的各文字，取出与图 12 示出的与各个汉字读音相对应的文字。

接着在 (c4) 中边参照从简繁体字对应装置 80 中取出的与上述对应的汉字和单词边删除不适当候补汉字和单词，便能得到图 13 示出的可能性高的汉字和单词。

在 (c5) 中利用最长一致法，转换为“他尚未使用软体”。

最后由输出单元 90 输出已经转换的文字串。

以上根据实施方式说明了本发明，本发明不限于上述实施方式，只要是在不变更其精神的范围内，就可以进行适当的改进并加以实施。例如：

30 (1)系统装置不分简体、繁体存储汉字和单词，而是用其涵义符号区分为简体和繁体并按照读音符号的顺序存储。

(2)读音的文字转换单元所使用的转换算法不局限于最长一致法，也可以采用考察未确认文字个数等的形态要素解析法，也可利用其他语法、使用频度等信息判断转换结果的正确性。具体地讲，就是用使用频度及 OCR 识别概率逼近候补文字（特别指定）或者进行排除。

5 (3)根据制造等情况，可将本发明必不可缺的构成要素（构成主要条件、事项、部分）看作是多个、或反过来将多个要素看作是一个，或者适当地对它们进行组合。

(4)在现有的字处理器和转换装置中读入必要的程序，再附加存储了装置的磁盘等，从而形成与本发明相同的结构。

10 (5)用价格高但具有更高速度的存储设备来存储使用频度高的文字，从而改进实施方式。

(6)输入单元通过用转换装置主体以及其它 L/E 型读取设备读入的并作为字节信息加以存储的磁盘来进行输入（读入）。输出设备则可输出在磁盘上代码化的文件。

15 (7)根据文件的领域，单词和文字的使用频度有很大不同。例如：在与动物有关的文件中“动”、“贝”、“马”、“猫（简体字）”、“鸟”等文字的使用频度高，而在与专利有关的文件中，“开发”等单词的使用频度高。因此，在读音的文字转换单元的转换算法采用或反映使用频度的情况下，可以使用按本装置使用者所输入的文件用途、内容范围而形成的频度表。或者用“开发”等特定的单词使用频度加以判断以后自动地改变参照频度表。

(8)在把用 OCR 输入的语句作为变更对象时，对装置中未记录的汉字判断为 OCR 读错误，因此要增加输出文字涵义的功能。

25 (9)对于近似形状的文字多、笔画多以及由其他经验事先知道 OCR 读取精度低的文字来说，应提高与包括该文字在内的单词、前后文字有关的权重，增加促使翻译者注意其涵义的功能。

正如以上说明的那样，如果采用本发明的中文简繁体字文件互换装置解决以前的问题，便可得到下述结果。

30 (1)由于利用文字的读音的信息，所以，可以在将读音转换为文字之时能有效地解决因单词未记在装置和相邻的两个单词转换问题（例如由于与“有一天”相应的读音符号串存在“有益”、“一天”等前后单词，所以会产生“有益天”的错误转换）以及同音异义字的选择问题（例如“上位”、

“尚未”、“上尉”三个同音异义字)等引起的问题。

通过利用语法、文法的读音信息,也可以减少变音字的错误转换,提高简体字文件和繁体字文件之间转换的正确率。而且可以减小装置的规模。

- 5 (2) 由于不使用 BIGRAM 信息,不影响语言资料库的内容就可以提高转换的正确性。不存在因语言资料收集困难而产生的转换不正确性。要想得到正确且符合每个领域的结果,就必须采用上述内容,以进一步提高精度。

- 10 (3) 作为中文字处理器等,在将中文文件从输入时的字音符号转换成汉字时以及在简体字和繁体字相互转换时均可使用同一系统装置。所以,在与中文字处理器等一道使用时,不需要设立其他的转换装置。进而可以减少编制装置的工时,也可节约经费。

- 15 (4) 词汇转换装置是一种语言记录的,所以只要使用这种装置,大陆、台湾就不需要在使用时繁杂地进行简体系统或繁体系统的转换,所以经济且省力。

所以,本发明的实用性非常高。

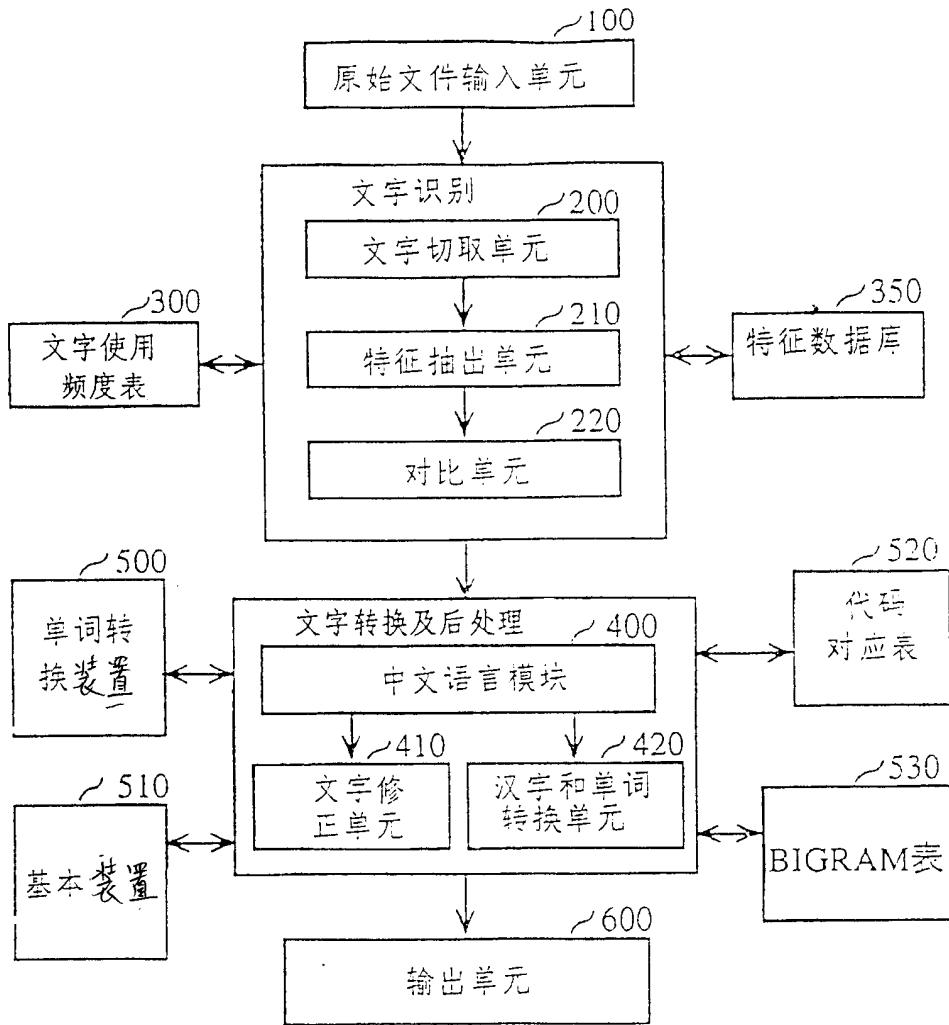


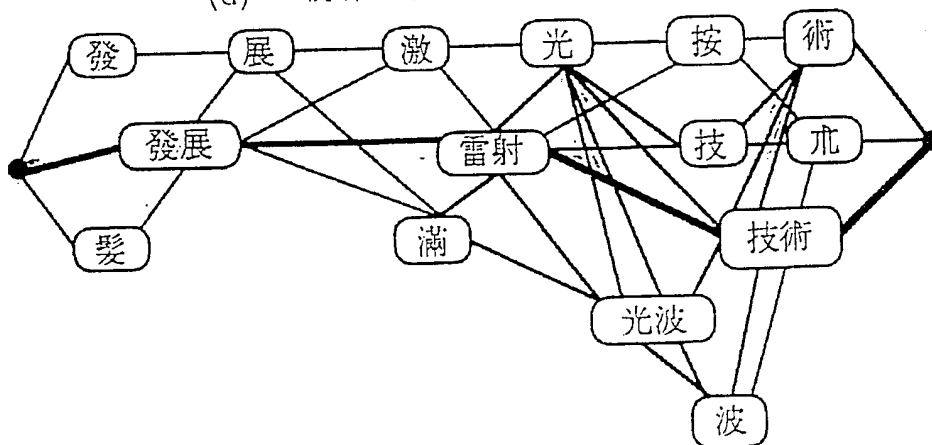
图 1

(a) 输入的简体字文件
发展激光技术

(b) 候补汉字
發展激光按術
髮 滿 技 朮
波

(c) 候补单词
發展，激光，光波

(d) 候补汉字和单词的文字网络



(e) 输出的转换结果
發展雷射技術

图 2

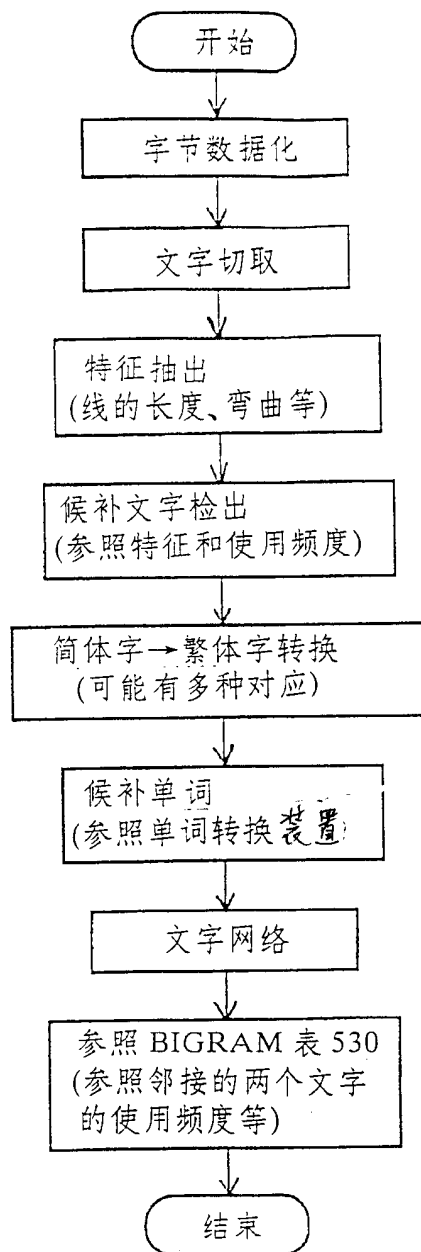


图 3

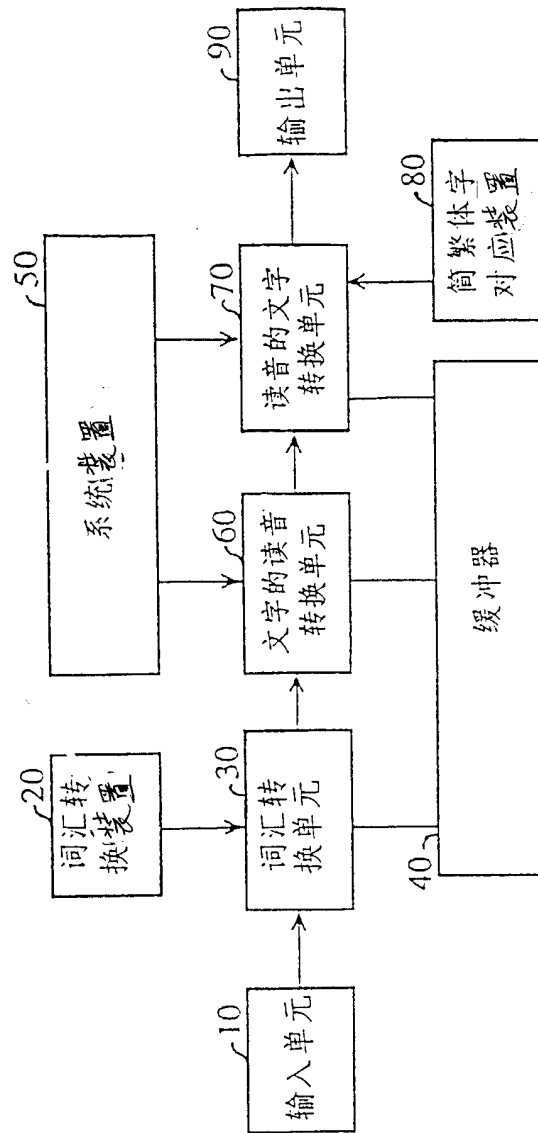


图 4

(a)

检索字	对应字
激光	雷射
打印	列印
软件	软体
硬件	硬体
⋮	⋮

(b)

检索字	对应字
列印	打印
雷射	激光
硬體	硬件
軟體	軟件
⋮	⋮

图 5

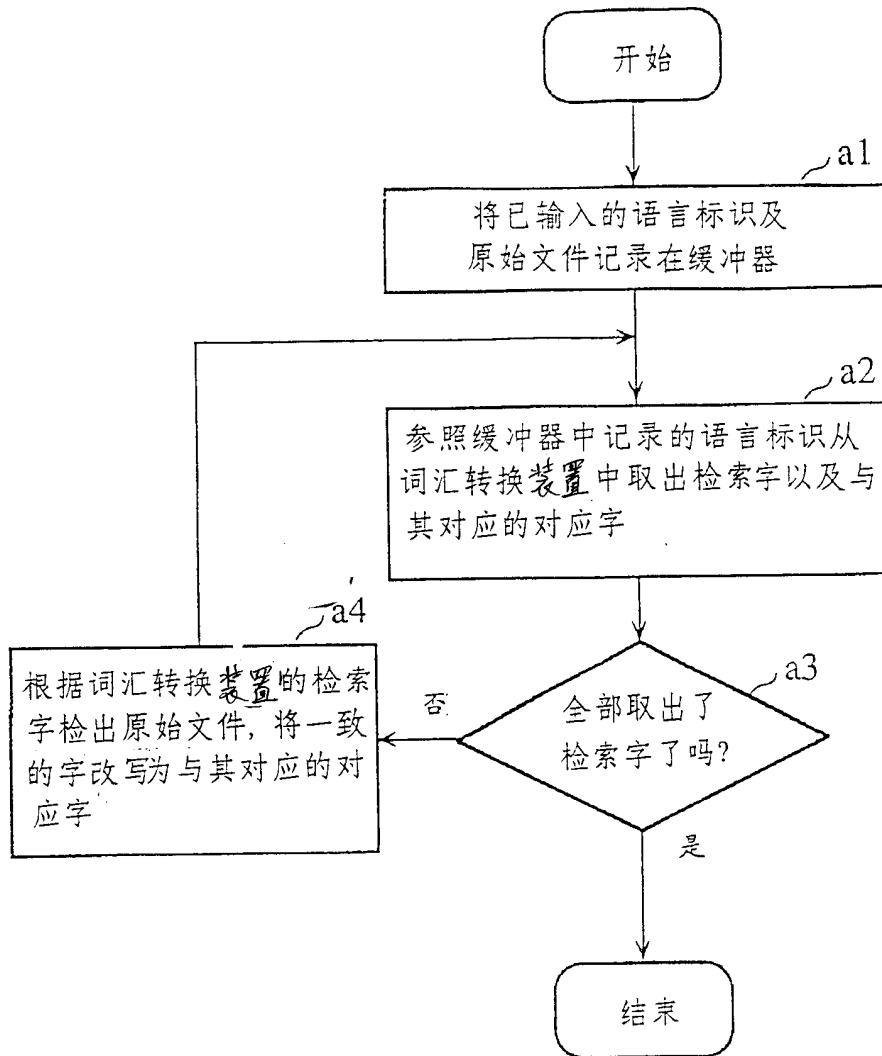


图 6

读音符号	候补汉字和单词
去Y 1(tal)	他，她，它，
尸尤4(shang4)	上，尚，
Xㄨ 4(uei4)	位，味，爲，
尸3(sh3)	使，史，駛，
ㄩㄥ 4(yueng4)	用，佣，
ㄩㄨㄢ 3(ruan3)	軟，阮，蠕，
去一3(ti3)	體，涕，
尸尤4 Xㄨ 4(shang4uei4)	上位，尚未，上尉
尸3 ㄩㄥ 4(sh3yueng4)	使用
ㄩㄨㄢ 3 去一3(ran3ti3)	軟體

图 7

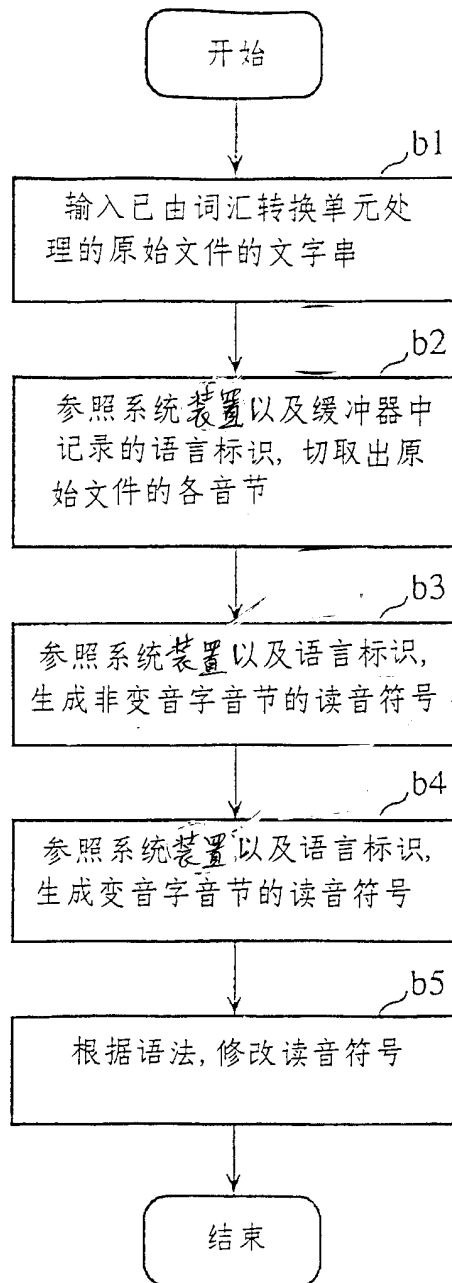


图 8

简体字	繁体字
汇	匯, 彙
贝	貝
并	並, 併
径	徑, 逕
尚	尙
未	未
⋮	

图 9

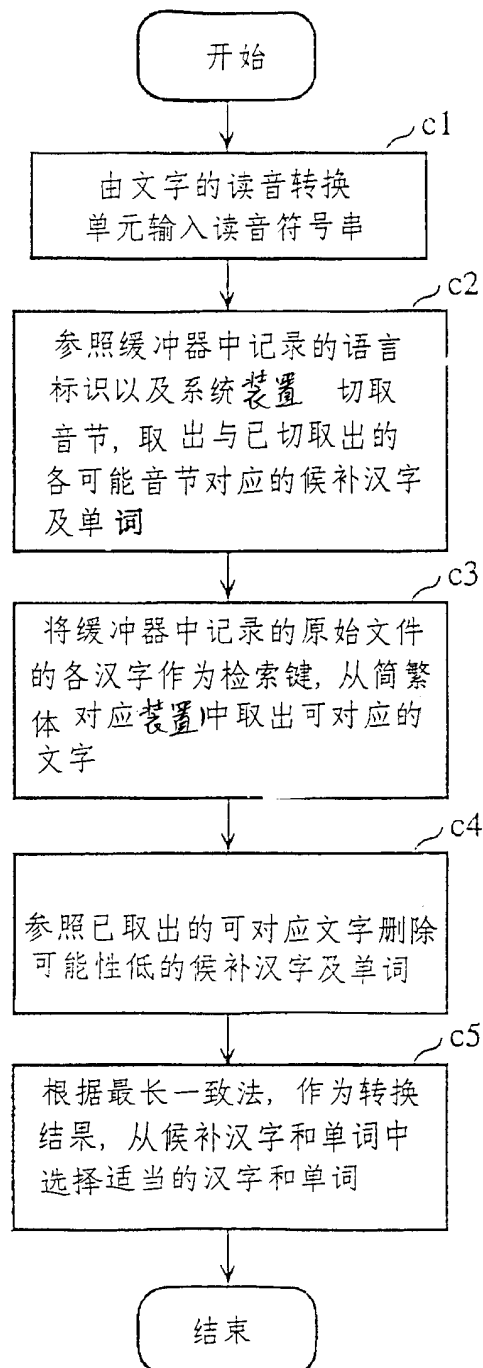


图 10

tal	shang4	uei4	sh3	yueng4	ruan3	ti3
他	上	未	史	用	軟	體
他	尚	胃	使	邕	阮	涕
:	:	:	:	:	:	:
:	:	:	:	:	:	:

尚未 使用 軟體

上尉

上位

图 11

tal	shang4	uei4	sh3	yueng4	ruan3	ti3
他	尚	未	使	用	軟	體
						体

图 12

tal	shang4	uei4	sh3	yueng4	ruan3	ti3
他	尚	未	使	用	軟	體
						體

尚未 使用 軟體

图 13