[19]中华人民共和国专利局

[11] 公开号 CN 1116343A

[12] 发明专利申请公开说明书

[21]申请号 94109394.8

[51]Int.Cl[6]

G06F 17/27

[43]公开日 1996 年 2 月 7 日

[22]申请日 94.8.5

[71]申请人 财团法人工业技术研究院

地址 中国台湾

[72]发明人 张照煌

[74]专利代理机构 中国专利代理(香港)有限公司

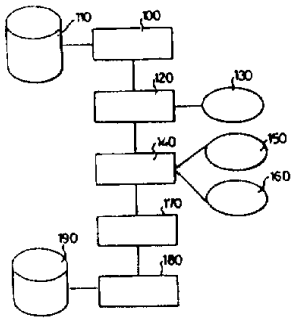
代理人 张志醒 王 岳

权利要求书 3 页 说明书 10 页 附图页数 4 页

[54]发明名称 中文错别字自动订正方法及装置

[57]摘要

一种自动检测并订正中文文书中错误字的方法及装置，先把原文书中的各中文字，以事先整理的含字形、字音、字义或与输入码相近字的综合近似字集进行代换，产生候选字串，再利用语言模型评分各候选字串，其中评分还包含"非原字扣分"，找出评分最高的候选字串，经与原文书中的中文字对照比较，即可自动检测出文书中的错别字所在，并提供对应的正确字，从而能有效地解决现有技术中误判率太高及未能提供正确字的缺点，极有实用价值。



权　　利　　要　　求　　书

1. 一种中文错别字自动检测订正方法，该方法是供电脑自动检测并订正中文文档中错别字的方法，其特征在于包括下列步骤:

综合近似字集代换步骤，将该文档中的文字以字形、字音、字义或与输入码相近字的综合近似字集的各文字予以代换，组合成多个候选字串;

语言模型评分步骤，利用一统计式语言模型对各候选字串给予评分，并找出评分最高的候选字串；及

错误字判断步骤，将上述评分最高的候选字串与所述文档中的文字逐字比对，并标示出其中相异的文字为错别字。

2. 如权利要求1 所述的方法，其特征在于，所述综合近似字集代换步骤中的综合近似字集由各文字包含原字的一个或多个字形、字音、字义或与输入码相近的文字组成。

3. 如权利要求2 所述的方法，其特征在于，所述综合近似字集中各文字的近似字分为多个等级。

4. 如权利要求1 所述的方法，其特征在于，所述综合近似字集代换步骤中，代换前先将所述文档中的文字根据标贴符号分成多个处理单元。

5. 如权利要求1 所述的方法，其特征在于，所述语言模型评分步骤对非原文档的文字评分予以扣分。

6. 如权利要求1 所述的方法，其特征在于，所述错别字判断步骤在标示错别字时，判断所述评分最高的候选字串中的对应文字为该错别字的正确字。

7. 一种中文错别字自动检测订正装置，该装置供电脑自动检测并

订正中文文档中的错别字，其特征在于它包括：

综合近似字集代换装置，用以将该文档中的文字代换成字形、字音、字义或与输入码相近字的文字，供组合成多个候选字串；

语言模型评分装置，用以对各候选字串给予评分，并找出评分最高的候选字串；及

错别字判断装置，用以逐字比对上述评分最高的候选字串与所述文档中的文字，并标示其中相异的文字为错别字。

8. 如权利要求7 所述的装置，其特征在于，所述综合近似字集代换装置包含一分割装置，用以在代换前先将所述文档中的文字根据标点符号分成多个处理单元。

9. 如权利要求7 所述的装置，其特征在于，所述综合近似字集代换装置包含：

综合近似字集数据库装置，内含中文字集中各文字包含原字的一个或多个字形、字音、字义或与输入码相近的文字；及

代换装置，将文字代换为综合近似字集装置内的近似字。

10. 如权利要求9所述的装置，其特征在于，所述综合近似字集装置中的综合近似字集数据库装置各文字的近似字分为多个等级。

11. 如权利要求7所述的装置，其特征在于，所述语言模型评分装置包含：

语言模型统计数据库，记录各语言单元的出现频率及语言单元之间的接续出现频率；

评分装置，根据一字串中所含的语言单元及语言模型统计数据库，评定该字串的分数；及

最高评分候选字串搜寻装置，决定最高评分的候选字串。

12. 如权利要求11 所述的装置，其特征在于，所述评分装置对非原文档的文字评分予以扣分。

13. 如权利要求11 所述的装置，其特征在于，所述语言模型评分装置的语言模型统计数据库包含一记录各词词类的中文词库。

14. 如权利要求11 所述的装置，其特征在于，所述语言模型评分装置以动态规划方式搜寻最高评分候选字串。

15. 如权利要求7所述的装置，其特征在于，所述错别字判断装置包含:

比对装置，逐字比对所述评分最高的候选字串与所述文档中的文字；及

标示装置，标示比对结果相异的文字为错别字。

16. 如权利要求7所述的装置，其特征在于，所述错别字判断装置在标示错别字时，判断所述评分最高的候选字串中的对应文字为该错别字的正确字。

说　　明　　书

中文错别字自动订正方法及装置

本发明有关于一种中文错别字自动订正方法及装置，特别是有关于利用综合近似字集代换及语言模型评分方式，使字形、字音、字义或与输入码相近的字集产生候选字串，并找出评分最高的候选字串，以便得到正确字的中文错别字自动订正方法及装置。

“错字”原指一中文字由于增减、改变笔画或偏旁误置所造成的讹字，“别字”则指不用某字而误用他字的情形，现今亦有人以“错字”一词涵盖“别字”的，以下统称为“错别字”。

错别字的多寡严重影响文书的质量，传统以人工一校再校的校稿订正，费时费力且常有漏校情形，如一般已经多校出版的报章杂志书籍，仍常见别字丛生。近年来由于电脑的普及，经输入电脑的文书虽免除了笔画错误造成的讹字，却也随之而产生由于输入过程造成的错误。所以利用电脑自动检测并订正错别字的需求实在非常迫切。

“检测错别字”指找出文书中错别字的所在，“订正错别字”则指找出该错别字的正确对应字。习知技术如已商品化的中文校稿系统仅有检测而没有订正的功能，本发明则同时具备检测和订正的功能。

电脑文书的错别字，不论来源于撰写创作过程或是输入编辑过程所产生的错误，均可分为以下四类或其中二类以上所共同造成：

(1) 同音或近音字，其发音相同或相近，

例1：　“行”　跡 可 疑（形）

例2：　按“步”　就 班（部）

(2) 字形相近字,

例3: 茶“壺”(壺)

例4: 桿“菌”(菌)

(3) 字义相近字,

例5: 既往不“究”(咎)

例6: 名不“符”實(副)

(4) 输入操作错误,即由于输入码相近造成的错别字或由于编辑操作错误产生的缺字、赘字或前后字互调,

例7: “糸”統(系,倉頡碼各爲VIF, HVIF)

例8: “坷”“坎”(坎坷),習慣“慣”()

根据这些加以整理分析,把一般人易犯错的字形、字音、字义或与输入码相近字进行汇集,使之成为综合近似字集数据库,用以代换原文书中的文字,产生候选字串,构成本发明的基础。

至于中文语言模型综合评分,含基底语言模型评分和“非原字扣分”。

其底语言模型评分可以利用习知的统计评分,如字接续表、词接续表、词间字接续表、词性接续表或词群接续表、或以词库为基础的词长词频评分,以几率值或分数值表现。“非原字扣分”则是对非原文字的近似字以分级或不分级的扣分。

利用语言模型综合评分,找出评分最高的候选字串,再与原文书中的中文字对照比较,即可自动检测出文书中的错别字所在并提供对应的正确字,极有实用价值。

现有技术中的台湾专利申请81104438号“中文错字自动侦测法及侦测装置”提出的中文错字自动检测法,主要包括两个步骤:(1) 假断词步骤,即参考一词库以找出无法形成复字词的和相邻字形成复字词的单字词,并将其取出;(2) 判断步骤,即根据各取出的单字词的

词频和前一字、后一字的接续强度来判断是否为正确字。该方法有两项缺点：(1)误判率太高，平均每四十个标示错误的字中只有一个真正的错字；(2)未能提供对应的正确字。

另有台湾专利申请80102492号“提高中文辨识率之错字更正法”和80107315号“文书辨识修正装置”，均为针对文字辨识装置产生的多候选字辨识结果做错字更正，与本发明无关。

又有美国专利如专利号为4,689,768(1987)，4,783,758(1988)，4,903,206(1988)，4,829,472(1989)，5,148,367(1992)的专利，均为针对如英文等西方语言的拼字检查订正，由于语言特性大不相同，因此是与本发明无关的技术。

与本发明有关的中文文书校稿系统，以往均借助断词后检测单字的词频和前后字接续强度的技术，故有误判率太高和未能提供对应的正确字等缺点和困难。本发明为克服这些缺点，提供了一种自动检测并订正中文错别字的方法及装置。

本发明的第一目的在于提供一种新颖的中文错字自动检测订正方法及装置。

本发明的第二目的在于提供检测出的错别字的正确对应字，以供订正。

本发明的再一目的在于降低错别字检测的误判率，提高自动校稿的效率。

为达到上述目的，本发明的中文错别字自动检测订正方法，是供电脑自动检测并订正中文文档中错别字的方法，该方法包括下列步骤：

综合近似字集代换步骤，将文档中的文字以字形、字音、字义或与输入码相近字的综合近似字集的各文字予以代换，组合成多个候选字串；

语言模型评分步骤，利用一统计式语言模型对各候选字串给予评

分，并找出评分最高的候选字串；及

错别字判断步骤，将该评分最高的候选字串与该文档中的文字逐字比对，并标示出其中相异的文字为错别字。

又，本发明的中文错别字自动检测订正装置，是供电脑自动检测并订正中文文档中错别字的装置，该装置包括：

综合近似字集代换装置，用以将文档中的文字代换成字形、字音、字义或与输入码相近字的文字，供组合成多个候选字串；

语言模型评分装置，用以对各候选字串给予评分，并找出评分最高的候选字串；及

错别字判断装置，用以逐字比对该评分最高的候选字串与该文档中的文字，并标示出其中相异的文字为错别字。

为清楚显示本发明的装置及方法，兹配合图示详细说明如下：

图1为本发明中文错别字检测订正装置实施例的方块图。

图2为本发明中文错别字检测订正方法的流程图。

图3为本发明实施例综合近似字集数据库的一部份。

图4为含有四个错别字的输入例句。

图5为该输入例句经近似字集代换后的结果。

图6为该输入例句经语言模型评分后的分数最高的五个候选字串。

图7为本发明实施例处理该例句后的输出结果。

本发明中文错别字检测订正装置实施例的组成如图1所示。

该装置主要包括：输入装置100、综合近似字集代换装置120、语言模型评分装置140、及错别字判断装置170、180。

输入装置100输入由使用者提供的中文文档110，并可包含一分割装置，用以在代换前先将该文档中的文字根据标点符号分成多个处理单元。

综合近似字集代换装置120，用以将该文档110中各文字代换成字

形、字音、字义或与输入码相近的文字，供组合成多个候选字串。该综合近似字集代换装置则包含：(a)综合近似字集数据库装置，内含中文字集中各文字包含原字的一个或多个字形、字音、字义或与输入码相近的文字，各文字的近似字还可分为多个等级；及(b)代换装置，将文字代换为综合近似字装置内的近似字。

语言模型评分装置140，用以对各候选字串给予评分，并找出评分最高的候选字串。该语言模型评分装置包含：(a)语言模型统计数据库，记录各语言单元的出现频率和语言单元之间的接续出现频率，其中还可包含一个记录各词词类的中文词库；(b)评分装置，根据一字串中所含的语言单元及语言模型统计数据库，评定该字串的分数，该评分装置对非原文档的文字予以扣分；及(c)最高评分候选字串搜寻装置，决定最高评分的候选字串，本实施例以动态规划方式搜寻最高评分候选字串。

错别字判断装置170、180，用以逐字比对该评分最高的候选字串与该文档中的文字，并标示出其中相异的文字为错别字。该错别字判断装置包含(a)比对装置170，逐字比对该评分最高的候选字串与该文档的文字；及(b)标示装置180，标示比对结果相异的文字为错别字，在标示错别字时，判断该评分最高的候选字串中的对应文字为该错别字的正确字，并将标示结果输出而成为一标示后文档190。

本发明中文错别字检测订正方法的处理流程如图2所示。此方法供电脑自动检测中文文档中的错别字，包括下列步骤：输入步骤200，输入一中文文档110，可先将该文档中的文字根据“，”、“。”、“？”、“！”、“；”、“：”等标点符号分成多个处理单元。对各处理单元的字串进行230至290各步骤，直至各处理单元均经处理后，结束步骤220；综合近似字集代换步骤230，将该文档110中的文字以字形(S)、字音(P)、字义(M)或与输入码(I)相近字的综合近似字

集130的各文字代换，组合成多个候选字串，其中综合近似字集由各文字包含原字的一个或多个字形、字音、字义或与输入码相近的文字组成，其中各文字的近似字还可分为多个等级；语言模型评分步骤240，利用一统计式语言模型250对各候选字串给予评分，其中语言模型评分对非原文档的文字评分予以扣分，利用Viterbi 动态规划方式搜寻最高评分候选字串(260)；及错别字判断步骤270，将该评分最高的候选字串与该文档中的文字逐字比对，并标示出其中相异的文字为错别字(280)，同时判断该评分最高的候选字串中的对应文字为该错别字的正确字；并将标示结果输出(290)成为一标示后文档(190)。

现举一例，说明本发明的实施过程。

假设“综合近似字集”为：

一 ：

人 ： 入S

力 ： 厲P 勵P 刀S 刃S

己 ： 已S 巳S 乙S

干 ： 甘P 乾P 千S

弋 ： 戈S

冶 ： 治S

究 ： 咎M 就P

利 ： 厲M 俐S 剁S 剎S 判S 力P

急 ： 岌P 疾M

糸 ： 系I

祇 ： 祗S 衹S 砥S 柢S 舐S 紙S 牴S 抵S

育 ： 育S 盲S

代换步骤以断句后的字串为处理单元，设原句为

S=C1,C2,...,Cn

把各中文字，经综合近似字集代换，产生候选字串：

P(i1,i2,...,in)=c1(i1),c2(i2),...,cn(in)

其中cj(ij)为含原字在内的第j个字的第ij个近似字,而1<=ij<=mj(ij=1表示使用原字)，

1<=j<=n，亦即共形成m1×m2×...×mn个候选字串。

利用语言模型评分各候选字串，其中进行评分并进行“非原字扣分”，找出评分最高的候选字串。

中文语言模型综合评分，包括基底语言模型评分和“非原字扣分”。

基底语言模型评分可以利用习知的统计评分，如字接续表、词接续表、词间字接续表、词性接续表或词群接续表、或以词库为基础的词长词频评分，以几率值或分数值表现。“非原字扣分”则是对非原文字的近似字予以分级或不分级的扣分。

本实施例所用的基底语言模型为词间字接续表和词库中的词频的联合评分，而“非原字扣分”以候选字串P(i1,i2,...,in)中用到的非原字近似字个数加权而得：

Penalty(P(i1,i2,...,in))=W×(ij !=1的个数)

FinalScore=BaseScore+Penalty

找出评分最高的候选字串的作法可采取穷举搜寻法或Viterbi 式动态规划搜寻法。

若找出评分最高的候选字串为P(k1,k2,...kn)，经与原文书中的中文字S=C1,C2,...Cn对照比较，即可自动检测出文书中的错别字所在，并提供对应的正确字。如cj(kj)不等于cj，则标示cj为错别字，并以cj(kj)为对应的正确字。

输出步骤：

输出各处理单元的处理结果，包括错别字标示和提供对应的正确字。

兹以一特定例句说明本发明的运作过程；

(1) 输入及断句步骤

“茶 壹 系 統 名 不 符 實 的 消 息 不 逕 而 走 。”

S =C1 C2 . . . C15

(2) 近似字集代换步骤

茶 壺 糸 統 名 不 符 實 的 消 息 不 逕 而 走

茶 壺 系 呂 副 銷 脛 爾

數 逍

计有2 ×2 ×2 ×2 ×3 ×3 ×2 ×2 =576个候选字串。

(3) 语言模型评分步骤

输出各处理单位的处理结果，包括错别字标示和提供对应的中文语言模型综合评分前五名候选字串，为：

名次	评分	候选字串
1	189-8=181	茶壺 系統 名不副實 的 消息 不脛而走
2	184-6=178	茶 壺 系統 名不副實 的 消息 不脛而走
3	182-6=176	茶壺 系統 名 不符 實 的 消息 不脛而走
4	177-4=173	茶 壺 系統 名 不符 實 的 消息 不脛而
5	181-10=171	茶壺 系統 名不副實 的 銷 息 不脛而走

(4) 对照比较步骤

原句： 茶壺系統名不符實的消息不逕而走

最高分: 茶壺系統名不副實的消息不脛而走

X X X X

(5)输出步骤

茶壺系統名不符實的消息不逕而走

壺系 副 脛

成功地检测并订正原句所有的四个错别字。

本发明功效的评估方法如下:

令 A＝输入文书的中文字总字数

B＝校稿方法标示错别字的字数

C＝校稿方法检测出并正确订正的字数

D＝校稿方法检测出为真实错别字的字数

E＝输入文书的真实错别字字数

则 标示率B-rate=B/A

准确率P-rate=D/B

检出率D-rate=D/E

订正率C-rate=C/E

现有中文校稿系统的指标: (见CCL Research Journal,1992.8)

B-rate=5.2% (太高)

P-rate=2.5% (太低)

D-rate=73.8%(可)

C-rate=0% (没有)

本发明的实施例经大量实验后结果如下:

(B,C,D为实施例指标, B′,D′为模拟习知中文校稿系统的指标)

测试资料	A	B	C	D	B′	D′	D和D′
国际政治	37114	13	6	6	2987	10	4

国际经济	87890	51	17	17	4721	15	12
国内政治	121863	73	34	34	8362	29	27
国际经济	110079	66	48	48	5526	47	45
	356946	203	105	105	21596	101	88

若D′为E的73.8%，则E=137，由此计算本发明的各项指标为

标示率 B-rate=B/A=203/356946=0.056%

准确率 P-rate=D/B=105/203 =51.72%

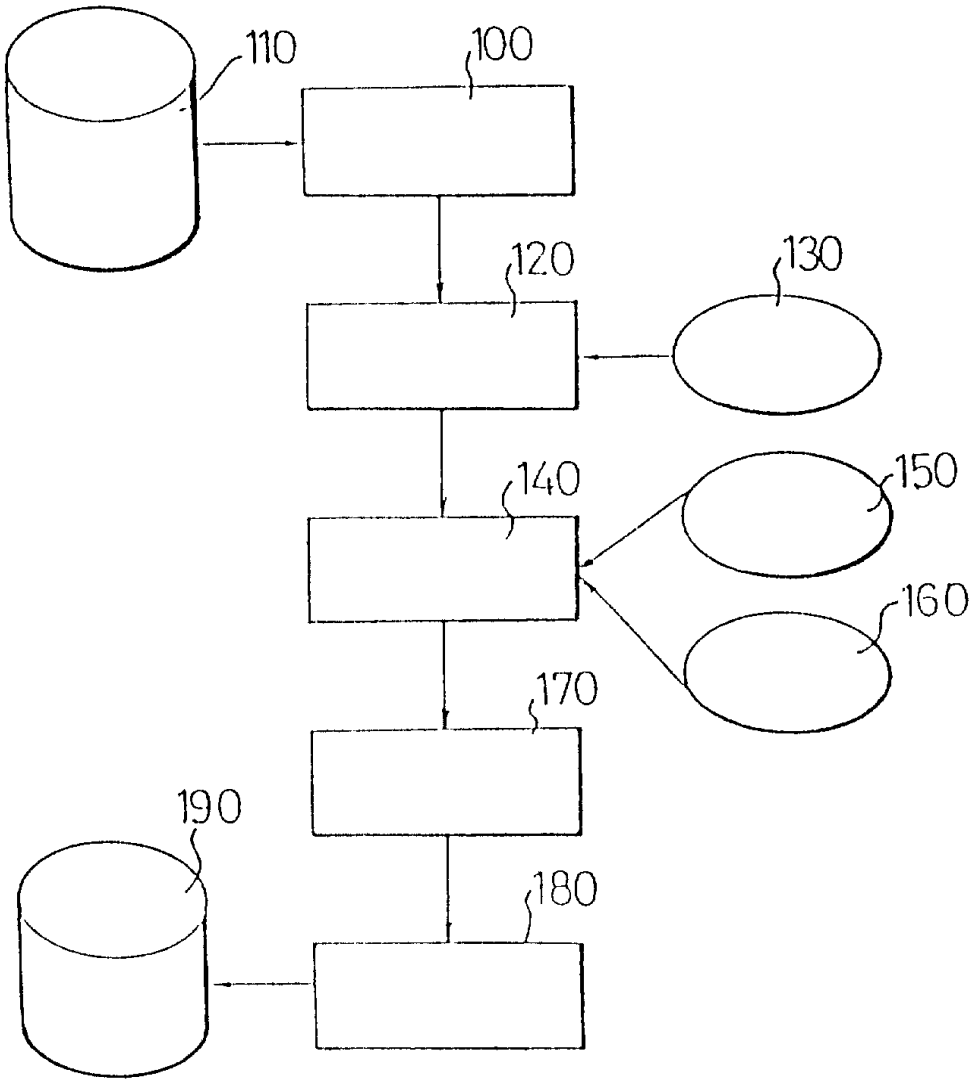
检出率 D-rate=D/E=105/137 =76.64%

订正率 C-rate=C/E=105/137 =76.64%

本发明的B-rate，P-rate，C-rate指标均远较习知技术为优，而D-rate大致相当，证明本发明极有实用价值。

说　明　书　附　图

图　1



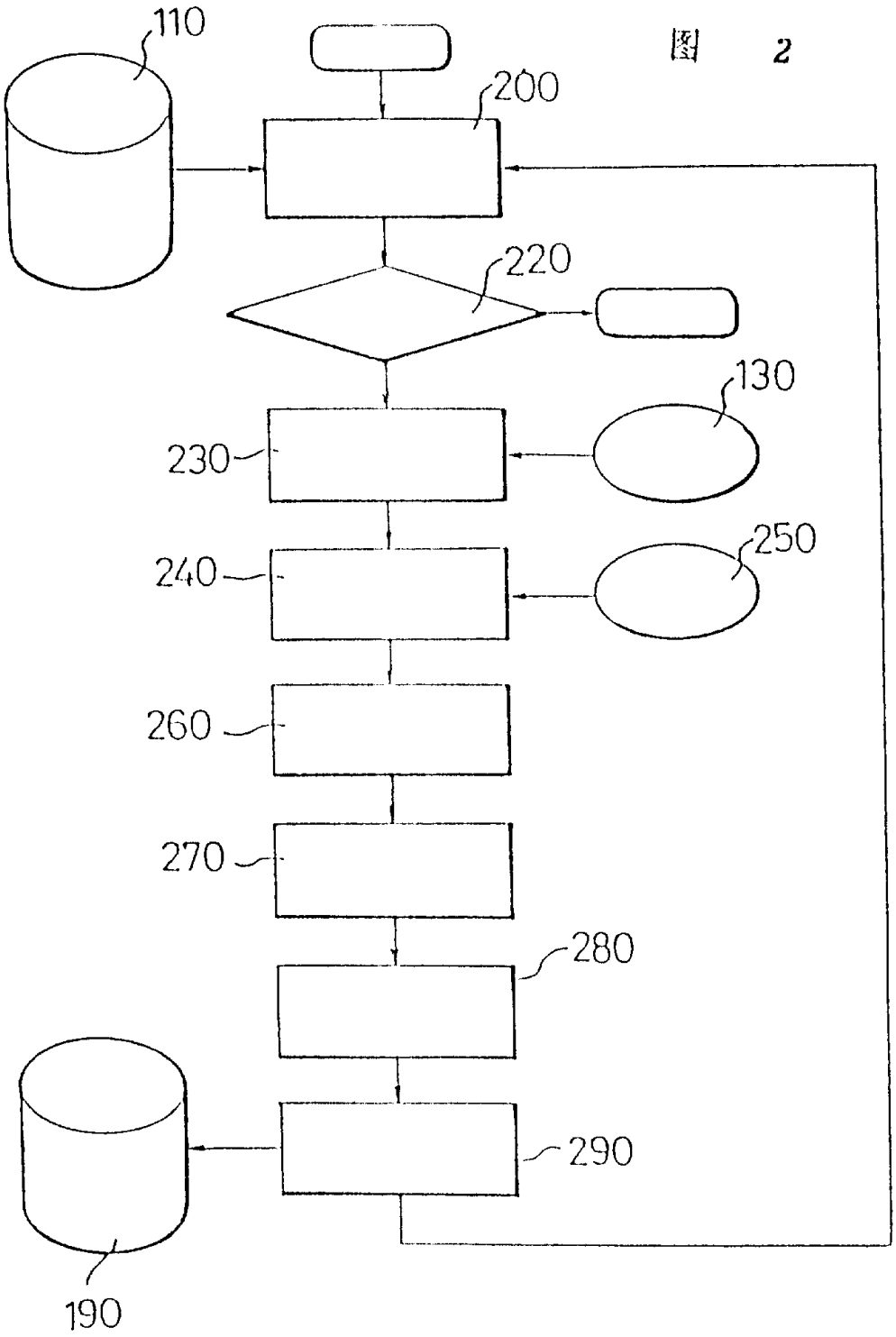


图 2

图　3

一 0

人 0　入 1

力 0　厲 1　励 2　刀 3　刃 3

已 0　已 1　巳 2　乙 5

干 0　甘 2　乾 3　千 5

戈 0　戈 2

冶 0　治 2

究 0　咎 2　就 3

利 0　厲 3　倒 3　剃 3　剎 3　判 5　力 6

急 0　岌 2　疾 3

糸 0　系 1

祇 0　祇 1　祇 1　砥 2　抵 2　舐 2　紙 3　牴 4　抵 6

育 0　育 2　盲 2

图　4

茶壺系統名不符實的消息不通而走。

图　5

茶壺系統名不符實的消息不通而走

茶壺系　呂　副　　銷　　脘痛

救　　道

图　6

181（189-8）　茶壶系统名不副实的消息不胫而走

178（184-6）　茶壶系统名不副实的消息不胫而走

176（182-6）　茶壶系统名不符实的消息不胫而走

173（177-4）　茶壶系统名不符实的消息不胫而走

171（181-10）　茶壶系统名不副实的销息不胫而走

图　7

茶壶系统名不符实的消息不逕而走

壶系　　副　　胫