[19]中华人民共和国专利局

[51]Int.Cl[6]
G06F 17/21

[12] 发明专利申请公开说明书

[21] 申请号 98104404.2

[43]公开日 1998 年 10 月 7 日

[11] 公开号 CN 1195142A

[22]申请日 98.2.12

[30]优先权

[32]97.3.28 [33]JP[31]77354/97

[71]申请人 松下电器产业株式会社

地址 日本大阪府

[72]发明人 郭俊桔

[74]专利代理机构 中国专利代理(香港)有限公司

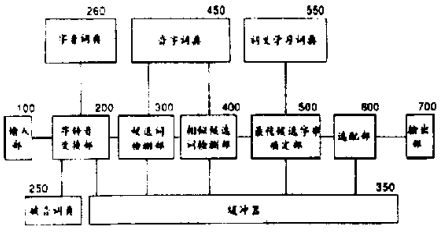
代理人 杨 凯 叶恺东

权利要求书 2 页 说明书 14 页 附图页数 16 页

[54]发明名称 汉语文档自动校正方法及其装置

[57]摘要

提供一种自动检测、修正汉语文档中的错别字、丢字的汉语文档自动校正方法及其装置。该装置包括：字转音变换部 200；候选词检测部 300；相似候选词检测部 400；最佳候选字串确定部 500；和选配部 600。



权　利　要　求　书

1. 一种用计算机自动校正电子化的汉语文档的汉语文档自动校正方法，其特征在于包括以下步骤:

词典制作步骤，在该步骤中预先编制以下各词典：将存储汉语中的各字的排列顺序的字顺序排列表和存储与其对应的读音符号的读音符号表同与上述读音符号表对应的全部候选词及其读音符号一起存储起来的破音字典；存储汉语中的文字符号和与其对应的错误读音符号及其它可能的读音符号的字音词典；以及存储汉语读音符号和与其对应的全部同音异义字、词及该全部同音异义字、词的使用频度加权和词义码的音字词典;

参照上述破音字典及字音词典，将由输入装置输入的原始文档中的字串变换成读音符号串的字转音变换步骤;

对在上述字转音变换步骤中获得的读音符号串分出音节，将上述分出的音节作为检索关键字，参照上述音字词典，检测出全部可能的候选词及其关连信息的候选词检测步骤;

利用掩蔽装置对连续汉字候选音节的相似位进行掩蔽，将掩蔽后的读音符号串作为检索关键字，参照上述音字词典，检测全部可能的候选词及其关连信息的相似候选词检测步骤;

将与原始文档中的字串对应的各候选词的开始位置、结束位置作为检索关键字，连接各候选词，形成有向网路，然后利用计算装置计算各候选词的相似度加权和词长加权，将使用频度加权、词长加权和原始文档相似度加权的累计最大值作为评价函数，利用动态规划法，取出最佳路径的最佳候选字串确定步骤;

以及对上述取出的最佳路径中的字串和原始文档字串进行选配，检测出不同的字后加以标记的选配步骤。

2. 一种用计算机自动校正电子化的汉语文档的汉语文档自动校正装置，其特征在于包括以下部分:

将存储汉语中的各字的排列顺序的字顺序排列表和存储与其对应的读音符号的读音符号表同与上述读音符号表对应的全部候选词及其读音符号一起存储起来的破音字典;

存储文字符号和与其对应的错误读音符号及其它可能的读音符号的字音词典;

存储读音符号和与其对应的全部同音异义字、词及该全部同音异义字、词的使用频度加权和词义码的音字词典;

参照上述破音字典及字音词典，将由输入装置输入的原始文档中的字串变换成读音符号串的字转音变换部;

对在上述字转音变换步骤中获得的读音符号串分出音节，将上述分出的音节作为检索关键字，参照上述音字词典，检测出全部可能的候选词及其关连信息的候选词检测部;

利用掩蔽装置对连续汉字候选音节的相似位进行掩蔽，将掩蔽后的读音符号串作为检索关键字，参照上述音字词典，检测出全部可能的候选词及其关连信息的相似候选词检测部;

将与原始文档中的字串对应的各候选词的开始位置、结束位置作为检索关键字，连接各候选词，形成有向网路，然后利用计算装置计算各候选词的相似度加权和词长加权，将使用频度加权、词长加权和原始文档相似度加权的累计最大值作为评价函数，利用动态规划法，取出最佳路径的最佳候选字串确定部;

以及对上述取出的最佳路径中的字串和原始文档字串进行选配，检测出不同的字后加以标记的选配部。

3. 根据权利要求 **2** 所述的汉语文档自动校正装置，其特征在于:

备有在存储器中存储着学习过的相邻接的后继词的词义码和前一词的词义码的组合的词义学习词典,

上述最佳候选字串确定装置将与原始文档中的字串对应的各候选词的开始位置、结束位置作为检索关键字，连接各候选词，形成有向网路后，利用计算装置计算各候选词的相似度加权和词长加权，参照上述词义学习词典，将使用频度加权、词长加权、原始文档相似度加权和词义相似度加权的累计最大值作为评价函数，利用动态规划法，取出最佳路径。

说　明　书

汉语文档自动校正方法及其装置

本发明涉及文档自动校正方法及其装置，特别是涉及自动地检测、修正汉语文档中的错字、丢字的汉语文档自动校正方法及其装置。

中国人写文章时容易写错字的原因如下所述:

（一）同音异义字或同音异形字

例如，[幼苗长得象笔]中的[得]，容易误写为“的”。另外，[清澈极了]中的[极]，容易误写为“急”。

（二）笔画错误

例如，[帽子]中的[帽]，一不小心就容易误写为“昌”等字。另外，“目”也容易误写为“日”。除此之外，笔画复杂的字、例如“龘”、“鬱”等也容易写错。

（三）字形相似

例如，[宰相]中的[宰]，其部首“宀”容易误写为“冖”。或者“辛”容易误写为“幸”。此外，[吵鬧]中的[鬧]，其部首“鬥”容易误写为“門”，[猫]的部首为[犭]也容易误写为部首“犬”。

（四）丢字

由于书写的速度太快，或不小心，容易造成丢字。例如将[辛辛苦苦]误写为[辛苦苦]。

（五）别字

使用别字。例如，[家庭]容易误写为[家廷]，[亭亭玉立]容易误写为[婷玉立]。

近十几年来，伴随电子计算机的进步和普及，汉语输入法也已创立了多种方案。汉语输入法按照编码方式大体可分为一般键盘输入和专用输入装置输入等。利用一般键盘的输入方式有如下几种:（一）按汉字的读音输入的方式，（二）按汉字的字形输入的方式，（三）字形和读音相结合的输入方式，（四）按文字码输入的方式，（五）按部首或笔画数输入的方式等。专用输入装置有例如专用的大键盘或 OCR（光学字符识别装置）等。

汉语输入法提供了一种能消除书写汉字时的困难的方法。可是，在用

计算机输入的汉语文档文件中，虽然能解决以往容易犯的错误中的例如笔画错误等问题，但其它错误仍然不可避免。一般而言，造成汉语文档文件中的错别字的原因可分为以下几种:

（一）不能正确地输入读音符号和字形的组合

一般可将汉语的读音符号分为声母、韵母、介音和声调。

声母: b、 p、 m、 ……

介音: i、 u、 yu、 iu、 ……

韵母: a、 o、 e、 ……

声调: 1（一声）、 2（二声）、 3（三声）、 4（四声）、 0（轻声）

例如，[形]的读音符号为[xing2]。

[字]的读音符号为[z4]。

其中，中国人容易混淆读音组如下:

声母部分: [sh]和[s]，或[q]和[x]等

介音部分: [i]和[yu]

韵母部分: [eng]和[en]，或[an]和[ang]等

声调部分: 很容易混淆。特别是对外国人来说，很难发出正确的声调。

例如，[兴趣]（ xing4qyu4 ）容易发成（ xing4qi4 ）的音，所以如果输入时，常被输入成"性器"。[学生]（ xyue2sheng1)和[写生](xie3sheng1)也容易互相误用。

另外，在字形输入的情况下，如果输入相似的字形组合或错误的组合，就不能得到正确的结果。例如，[日]和[曰]、[受]和[爱]等字形组合码极其相似。

（二）同音异义字

选择错误的同音异义字、词。例如，[同音异义字]容易选择成[同音意义字]、或[同音异议字]。

（三）参照词典的错误

无论采用哪一种输入法，都要利用参照词典进行变换。如果该参照词典的内容有错误，那么输入结果当然也就会错。例如在参照词典中，如果将[形影不离]这一成语登录为[行影不离]，在输入了前一个读音符号后，其变换结果必然是[行影不离]这样的误变换。

（四）输入操作的错误

通常在编辑文档时，都要利用各种文档编辑软件。可是，在执行[插入]或[删除]等功能时，如果不特别注意地进行操作，很可能在文档中造成多字或丢字的现象。

在汉语文档文件中，错字会极大地影响文档的质量。所以，如何能有效地检测、并自动地修正汉语文档文件中的错误已成为重要课题。

作为现有的汉语文档自动校正方法及其装置，有例如中华民国专利公告第 260772 号所述的方法及装置。图 17 就是该现有例的结构图。在该图中，100 是输入欲处理的汉语文档的输入装置。110 是存储欲处理的汉语文档的汉语文档文件。120 是参照综合相似字集，将输入的汉语文档中的每个字变换成相似字，根据变换后的相似字组合成多个候选字串的综合相似字形变换装置。130 是存储汉字的字形、字音、字义或输入码相似字的综合相似字集。以下所示就是该例（S：字形相似，P：字音相似，M：字义相似，I：输入码相似）。

人：入 S

力：歷 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

140 是对各候选字串进行评价，根据评价过的各候选字串，检测评价最高的候选字串的语言模型评价装置。150 是评价装置，它由(a)语言模型统计数据库和(b)评价装置构成。(a)语言模型统计数据库包含记录各语言单位的使用频度、语言单位之间的连续使用频度，且记录各词的频度的汉语知识库。(b)评价装置对于一字串来说，根据语言单位及语言模型统计数据库，对该字串加分。此后，对该原始文档文件中没有的字减分。160 是用动态规划法检测评价最高的候选字串的最高评价候选字串检测装置。170 是以逐字选配方式对照该最高评价候选字串和该文档文件中的字串，将不同的字作为错字并进行显示的错别字判断装置。180 是将显示过的字串输出给显示后的文档文件的显示结果输出装置。190 是存储所显示的字串的显示后文档文件。

以下说明该现有例的工作情况。

利用输入装置 100 从汉语文档文件 110 输入欲处理的汉语文档。根

据标点符号的位置，将上述输入的汉语文档分成若干个处理单位，输入综合相似字形变换装置 120 。在综合相似字形变换装置 120 中，按照各处理单位参照综合相似字集 130 ，取出全部字形、字音、字义或输入码相似的字，组合成多个候选字串，然后输入语言模型评价装置 140 。在语言模型评价装置 140 中，根据评价装置 150 中的统计的语言模型，评价各候选字串，根据语言模型评价情况，对原始文档文件中没有的字减分。在最高评价候选字串检测装置 160 中，利用动态规划法检测最高评价候选字串，之后输入到错别字判断装置 170 。在错别字判断装置 170 中，依次对照（选配）最高评价候选字串和输入的原始文档文件，将不同的字作为错别字显示，输入到显示结果输出装置 180 。显示结果输出装置 180 将显示字串输出给显示后文档文件 190 。

可是，与上述现有的技术有关的方法及其装置存在以下问题:

（ 1 ）不能检测、修正综合相似字集中未登录的字。因此，为了制作、保存知识库，就需要花费大量的人力、物力或经费。

（ 2 ）在语言模型评价装置中，只考虑了各词的出现频度和语言单位之间的连续使用频度，而没有利用词义信息，所以检测率和修正率不高。

（ 3 ）不能消除汉语文档中的丢字、多字、字的顺序错误等问题。

为了达到上述目的，本发明的第一方面是关于用计算机自动校正电子化的汉语文档的汉语文档自动校正方法，其特征在于包括以下步骤:

词典制作步骤，在该步骤中预先编制以下各词典: 将存储汉语中的各字的排列顺序的字顺序排列表和存储与其对应的读音符号的读音符号表同与上述读音符号表对应的全部候选词及其读音符号一起存储起来的破音字典; 存储汉语中的文字符号和与其对应的错误读音符号及其它可能的读音符号的字音词典; 以及存储汉语读音符号和与其对应的全部同音异义字、词及该全部同音异义字、词的使用频度加权和词义码的音字词典;

参照上述破音字典及字音词典，将由输入装置输入的原始文档中的字串变换成读音符号串的字转音变换步骤;

对在上述字转音变换步骤中获得的读音符号串分出音节，将上述分出的音节作为检索关键字，参照上述音字词典，检测出全部可能的候选词及其关连信息的候选词检测步骤;

利用掩蔽装置对连续汉字候选音节的相似位进行掩蔽，将掩蔽后的

读音符号串作为检索关键字，参照上述音字词典，检测全部可能的候选词及其关连信息的相似候选词检测步骤；

将与原始文档中的字串对应的各候选词的开始位置、结束位置作为检索关键字，连接各候选词，形成有向网路，然后利用计算装置计算各候选词的相似度加权和词长加权，将使用频度加权、词长加权和原始文档相似度加权的累计最大值作为评价函数，利用动态规划法，取出最佳路径的最佳候选字串确定步骤；

以及对上述取出的最佳路径中的字串和原始文档字串进行选配，检测出不同的字后加以标记的选配步骤。

本发明的第二方面是关于用计算机自动校正电子化的汉语文档的汉语文档自动校正装置，其特征在于包括以下部分：

将存储汉语中的各字的排列顺序的字顺序排列表和存储与其对应的读音符号的读音符号表同与上述读音符号表对应的全部候选词及其读音符号一起存储起来的破音字典；

存储文字符号和与其对应的错误读音符号及其它可能的读音符号的字音词典；

存储读音符号和与其对应的全部同音异义字、词及该全部同音异义字、词的使用频度加权和词义码的音字词典；

参照上述破音字典及字音词典，将由输入装置输入的原始文档中的字串变换成读音符号串的字转音变换部；

对在上述字转音变换步骤中获得的读音符号串分出音节，将上述分出的音节作为检索关键字，参照上述音字词典，检测全部可能的候选词及其关连信息的候选词检测部；

利用掩蔽装置对连续汉字候选音节的相似位进行掩蔽，将掩蔽后的读音符号串作为检索关键字，参照上述音字词典，检测全部可能的候选词及其关连信息的相似候选词检测部；

将与原始文档中的字串对应的各候选词的开始位置、结束位置作为检索关键字，连接各候选词，形成有向网路，然后利用计算装置计算各候选词的相似度加权和词长加权，将使用频度加权、词长加权和原始文档相似度加权的累计最大值作为评价函数，利用动态规划法，取出最佳路径的最佳候选字串确定部；

以及对上述取出的最佳路径中的字串和原始文档字串进行选配，检测

不同的字后加以标记的选配部。

本发明的第三方面是第二方面所述的汉语文档自动校正装置，其特征在于：设有在存储器中存储着学习过的相邻接的后继词的词义码和前一词的词义码的组合的词义学习词典，上述最佳候选字串确定装置将与原始文档中的字串对应的各候选词的开始位置、结束位置作为检索关键字，连接各候选词，形成有向网路后，利用计算装置计算各候选词的相似度加权和词长加权，参照上述词义学习词典，将使用频度加权、词长加权、原始文档相似度加权和词义相似度加权的累计最大值作为评价函数，利用动态规划法，取出最佳路径。

由于如上构成，本发明的第一方面所述的用计算机自动校正电子化的汉语文档的汉语文档自动校正方法，是在词典制作步骤中编制将存储汉语中的各字的排列顺序的字顺序排列表和存储与其对应的读音符号的读音符号表同与上述读音符号表对应的全部候选词及其读音符号一起存储起来的破音字典；另外编制存储汉语中的文字符号和与其对应的错误读音符号及其它可能的读音符号的字音词典；还编制存储汉语读音符号和与其对应的全部同音异义字、词及该全部同音异义字、词的使用频度加权和词义码的音字词典。在字转音变换步骤中，参照上述破音字典及字音词典，将由输入装置输入的原始文档中的字串变换成读音符号串。在候选词检测步骤中，对在上述字转音变换步骤中获得的读音符号串分出音节，将上述分出的音节作为检索关键字，参照上述音字词典，检测全部可能的候选词及其关连信息。在相似候选词检测步骤中，利用掩蔽装置对连续汉字候选音节的相似位进行掩蔽，将掩蔽后的读音符号串作为检索关键字，参照上述音字词典，检测全部可能的候选词及其关连信息。在最佳候选字串确定步骤中，将与原始文档中的字串对应的各候选词的开始位置、结束位置作为检索关键字，连接各候选词，形成有向网路，然后利用计算装置计算各候选词的相似度加权和词长加权，将使用频度加权、词长加权和原始文档相似度加权的累计最大值作为评价函数，利用动态规划法，取出最佳路径。在选配步骤中，对上述取出的最佳路径中的字串和原始文档字串进行选配，检测不同的字后加以标记。

在本发明的第二方面所述的用计算机自动校正电子化的汉语文档的汉语文档自动校正装置中，破音字典将存储汉语中的各字的排列顺序的字顺序排列表和存储与其对应的读音符号的读音符号表同与上述读音符

号表对应的全部候选词及其读音符号一起存储起来。字音词典存储文字符号和与其对应的错误读音符号及其它可能的读音符号。音字词典存储读音符号和与其对应的全部同音异义字、词及该全部同音异义字、词的使用频度加权和词义码。字转音变换部参照上述破音字典及字音词典，将由输入装置输入的原始文档中的字串变换成读音符号串。候选词检测部对在上述字转音变换步骤中获得的读音符号串分出音节，将上述分出的音节作为检索关键字，参照上述音字词典，检测全部可能的候选词及其关连信息。相似候选词检测部利用掩蔽装置对连续汉字候选音节的相似位进行掩蔽，将掩蔽后的读音符号串作为检索关键字，参照上述音字词典，检测全部可能的候选词及其关连信息。最佳候选字串确定部将与原始文档中的字串对应的各候选词的开始位置、结束位置作为检索关键字，连接各候选词，形成有向网路，然后利用计算装置计算各候选词的相似度加权和词长加权，将使用频度加权、词长加权和原始文档相似度加权的累计最大值作为评价函数，利用动态规划法，取出最佳路径。选配部对上述取出的最佳路径中的字串和原始文档字串进行选配，检测不同的字后加以标记。

在本发明的第三方面中，词义学习词典存储着学习过的相邻接的后继词的词义码和前一词的词义码的组合。上述最佳候选字串确定装置将与原始文档中的字串对应的各候选词的开始位置、结束位置作为检索关键字，连接各候选词，形成有向网路后，利用计算装置计算各候选词的相似度加权和词长加权，参照上述词义学习词典，将使用频度加权、词长加权、原始文档相似度加权和词义相似度加权的累计最大值作为评价函数，利用动态规划法，取出最佳路径。

图 1 是表示 2 字节的汉语中的汉字读音结构的结构图。

图 2 是举例说明相似的音韵要素之间的位图之间的距离的说明图。

图 3 是本发明的一实施例的结构图。

图 4 是上述实施例中的字转音变换部的工作流程图。

图 5 是上述实施例中的候选词检测部的工作流程图。

图 6 是上述实施例中的相似候选词检测部的工作流程图。

图 7 是上述实施例中的最佳候选字串确定部的工作流程图。

图 8 是上述实施例中的选配部的工作流程图。

图 9 是上述实施例中的破音字典的示意图。

图 10 是上述实施例中的字音词典中的数据结构的示意图。

图 11 是上述实施例中的音字词典中的数据结构的示意图。

图 12 是上述实施例中的词义学习词典中的数据结构的示意图。

图 13 是上述实施例中的呈层次的词义分类的示意图。

图 14 是根据具体例说明上述实施例中的处理内容的说明图。

图 15 是继图 14 的说明图。

图 16 是继图 14 的说明图。

图 17 是现有的汉语文档自动校正方法及其装置的结构图。

“词义”是形态要素本身的意义（或称词义码）。在以下所述的实施例中，采用了由角川书店出版的相似词词典（1985 年）中记载的词义分类方法。该词义分类方法是用大类（第一位）、中类（第二位）、小类（第三位）、细类（第四位）这种 16 进制的数字表示一个形态要素的全部分类信息。另外，这里之所以采用 16 进制的数字，是因为计算机中广泛地采用 16 进制（2 字节）的数字，而且如果是 16 进制，那么用 1 位就能充分地对应各种分类。该相似词词典将全部汉字、单词等分成“自然”、“性状”、“变动”、“行动”、“心情”、“人物”、“成功”、“社会”、“学术”、“物品”等十大类，另外将各大类分成十个中类，各中类、小类也用同样的方法分成更细的类。在本实施例中，在该四位数字之前加 s，如下表示：

s0　　（属于“自然”类）

s02　　（属于“自然”类的“气象”）

s028　　（属于“自然”类的“风”）

s028a　　（属于“自然”类的“强弱”）

这样的层次分类码举例示于图 13。在这样的层次分类码中，高位的词义码的意义范围比低位的广。就是说越是低位的词义码的意义范围越窄。因此，可以配合实际的需要来利用词义码，由于预先没有必要，所以也不需要一一登录，能节省存储器。另外，由于该词义码用数字表示，所以在通过数学运算、例如用集合逻辑积处理词义码的情况下，有可能获得由词义码生成的更有价值的信息。另外，关于词义码的详细说明发表于特开平 3-202954 号公报中，故这里从略。

汉语中的汉字的读音的种类约有 1300 种，如果对其编码的话，最多

使用2字节（以字节为单位的情况）就够了，但其中声母（子音）有22个，介音有3个，韵母（母音）有14个，声调有5个。2字节的汉语中的汉字读音的结构示于图1。第一字节含有声母（位2~位6）和介音（位0~1），第二字节含有声调（位4~位6）和韵母（位0~3）。因此，例如将第一字节的介音区掩蔽起来，使用逻辑积运算装置就能检测具有相同的声母、韵母和声调的字。

为了利用掩蔽方法处理各区的相似音，则使各区中的相似的音韵要素之间的位图之间的距离为1。将该例示于图2。

关于汉语读音压缩码和相似位配置的详细说明请参照特公平7-60433号“汉字变换装置”。在欲处理由于编辑上的错误而造成的多字、丢字、字的顺序错误等问题的情况下，本实施例中的掩蔽方法除了能进行上述的位的掩蔽之外，还能将字完全掩蔽。以“ting2* yu4 1i4”为例，将“ting2 yu4 1i4”或“*ting2 yu4 1i4”（*表示被掩蔽的字。即，任何字都可以）作为检索关键字，通过参照音字词典，能检测“亭亭玉立”。

另外，如果参照特公平7-60434号“汉字变换装置”，可知在将读音符号串（表音符号串）变换为字串（汉字串）时，字数多的单词（单词的长度，这里称作词长）成为一个重要的评价要素。因此，在本实施例中，词长加权也作为一个评价函数。其计算式如下所示。例如在候选词为“大家”的情况下，其词长加权为（2－1)*2=2。

词长加权=（候选词的字数－1）*2

另外，为了利用原始文档中的字信息，有效地取出最佳路径，在本实施例中，将原始文档相似度加权作为一个评价函数。其计算式如下所示。

原始文档相似度加权=(选配原始文档中的字和与其对应的候选词而具有相同的字的个数）/候选词的字数

例如，与原始文档中的字“亭玉立”对应的候选词为“亭亭玉立”时，该候选词的原始文档相似度加权为3/4（0.75）。

在本实施例中，还导入前一词、后继词的词义信息。例如，如图12所示，该单词的词义是根据作了标记后的大型词库（コパス），参照前后词的词义，自动学习后获得的。或者是根据对不同的区作了标记后的文档集学习获得的。由于采用层次定义方式，所以通过集合逻辑集的运算，进行前后词的词义相似度计算，可获得词义码。例如，词义码[7140]

和[714a]的集合逻辑集的运算结果为[714]。这时，由于三个码一致，所以词义码相似度为 3/4 。另外，当全部码一致时，词义相似度为 1 ，在两个码一致的情况下，词义相似度为 2/4 ，在一个码一致的情况下，词义相似度为 1/4，完全不一致的情况下为 0 。

图 3 是本发明的一实施例的结构图。在该图中， 250 是将存储汉语中的各字的排列顺序的字顺序排列表和存储与其对应的读音符号的读音符号表同与上述读音符号表对应的全部候选词及其读音符号一起存储起来的破音字典。破音字典的示意图示于图 9 。 260 是存储文字符号和与其对应的错误读音符号及其它可能的读音符号的字音词典。字音词典的数据结构示意图示于图 10 。 450 是存储读音符号和与其对应的全部同音异义字、词及该全部同音异义字、词的使用频度加权（长期学习）和词义码的音字词典。音字词典的数据结构示意图示于图 11 。 550 是存储着学习过的相邻接的后继词的词义码和前一词的词义码的组合的词义学习词典。词义学习词典的示意图示于图 12 。 350 是暂时记录中间处理数据的缓冲器。 100 是例如硬盘、键盘等输入装置，是输入原始文档的输入部。 200 是参照上述破音字典 250 及字音词典 260 ，将由输入装置输入的原始文档中的字串变换成读音符号串的字转音变换部。 300 是对由上述字转音变换部 200 获得的读音符号串分出音节，将上述分出的音节作为检索关键字，参照上述音字词典 450 ，检测全部可能的候选词及其关连信息的候选词检测部。 400 是利用掩蔽装置对连续汉字候选音节的相似位进行掩蔽，将掩蔽后的读音符号串作为检索关键字，参照上述音字词典 450 ，检测全部可能的候选词及其关连信息的相似候选词检测部。 500 是将与原始文档中的字串对应的各候选词的开始位置、结束位置作为检索关键字，连接各候选词，形成有向网路，然后利用计算装置计算各候选词的相似度加权和词长加权，参照上述词义学习词典 550 ，将使用频度加权 + 词长加权 + 原始文档相似度加权 + 词义相似度加权的累计最大值作为评价函数，利用动态规划法，取出最佳路径的最佳候选字串确定部。 600 是对上述取出的最佳路径中的字串和原始文档字串进行选配，检测不同的字后加以标记的选配部。 700 是输出上述最佳字串和标记后的原始文档的字串的输出部。

本实施例中的字转音变换部 200 的工作流程示于图 4 。

以下，参照该图说明其工作情况。

（S201） 由输入部100将原始文档中的字输入后，记录在缓冲器350中。

（S202） 参照字音词典260，按每一音节分出原始文档中的字，然后进入（S203）的处理。

（S203） 分别取出记录在缓冲器350中的各音节，参照字音词典260，将非破音字变换成读音符号后记录在缓冲器350中。

（S204） 参照破音字典250，将记录在缓冲器350中的有破音字的字变换成适合于破音字的读音符号。

（S205） 参照缓冲器350中的原始文档中的字，根据汉语语法修正缓冲器350中的各字的读音符号。例如，“妈”的读音为“mal”，但“妈妈”中的第二个“妈”的声调不用1声（四声中的最高声）读，而应该用轻声（由于音节连续而失去固有的声调，轻轻地发音）读成“mao”，所以修正第二个“妈”的读音符号。

至此，字转音变换部200的处理结束。

图5示出了本实施例中的候选词检测部300的工作流程。

以下，参照该图说明其工作情况。

（S301） 输入由字转音变换部200获得的原始文档的读音符号。参照音字词典450，将读音符号分成可能成为音节的全部音节，然后进入（S302）的处理。

（S302） 将分出的音节作为检索关键字，从音字词典450中取出全部候选词及其使用频度加权、以及词义码。

（S303） 将候选词及其关连信息记录在缓冲器350中后，结束处理。

图6示出了本实施例中的相似候选词检测部400的工作流程。

以下，参照该图说明其工作情况。

（S401） 由候选词检测部300输入读音符号串和全部候选词及其关连信息，然后进入（S402）。

（S402） 判断是否有未处理的连续单音节候选。在还有未处理的连续单音节候选的情况下，进入（S403）。在没有未处理的连续单音节候选的情况下，结束相似候选词检测部400的处理。

（S403） 由连续单音节的候选读音和相似音掩蔽装置参照音字词典450，取出相似词及其关连信息，然后进入（S404）。

（S404） 参照读音符号串，对上述取出的相似候选词计算各相似候选词的开始位置、结束位置，然后返回（S402）的处理。

图7示出了本实施例中的最佳候选字串确定部500的工作流程。以下，参照该图说明其工作情况。

（S501） 而由相似候选词检测部400输入相似候选词及其关连信息，然后进入（S502）。

（S502） 从缓冲器350取出候选词及其关连信息，然后进入（S503）的处理。

（S503） 将各候选词的开始、结束位置作为检索关键字，作成候选词的有向网路。

（S504） 从缓冲器350取出原始文档的字串，将各候选词的开始位置、结束位置作为检索关键字，计算原始文档的相似度加权、词长加权，然后进入（S505）的处理。

（S505） 将使用频度加权＋词长加权＋原始文档相似度加权＋词义相似度加权的累计最大值作为评价函数，利用动态规划法，取出最佳路径。然后进入（S506）。

（S5056） 取出最佳路径中的候选词，然后将其输出。

图8示出了本实施例中的参照选配部600的工作流程。

以下，参照该图说明其工作情况。

（S601） 由最佳候选字串确定部500输入最佳路径中的字串A，然后进入（S602）。

（S602） 取出缓冲器350中记录的原始文档的字串B，然后进入（S603）的处理。

（S603） 由掩蔽装置对字串A和字串B进行掩蔽后，再对原始文档中的错别字、词进行掩蔽，然后进入（S604）的处理。

（S604） 上述掩蔽后的原始文档的字串和最佳路径中的字串被送给输出部700。

以下，以输入了“多词资料库系统”的情况为例，具体地说明如上构成的本实施例的工作情况。如果从输入部100输入了称为“多词资料库系统”的原始文档，则字转音变换部200参照字音词典260和破音字典250，将上述输入的原始文档变换成如下所示的读音符号“duo1yu3z11iao4ku4xi4tueng3”，然后记录在缓冲器350中。其次，

候选词检测部将上面所示的读音符号分成可能成为音节的全部音节。如图 14（3）所示，将上述分出的音节作为检测关键字，参照音字词典 450，检测出全部可能的候选词及其关连信息。然后输入到相似候选词检测部 400。由于只在“duo1 yu3”音节中有候选字，所以由上述的音节压缩装置和掩蔽装置参照音字词典 450，对上述两个音节检测图 14（4）所示的相似候选词及其关连信息，然后进入最佳候选字串确定部 500 的处理。最佳候选字串确定部 500 首先将与原始文档的字串对应的开始位置、结束位置作为检索关键字，连接各候选词，作成图 15（5）所示的有向网路。然后，参照词义学习词典 550，将使用频度加权 + 词长加权 + 原始文档相似度加权 + 词义相似度加权的累计最大值作为评价函数，利用动态规划法，能检测出图 15（5）所示的最佳路径。由此进入选配部 600 的处理。选配部 600 取出缓冲器 350 中记录的原始文档的字串。由选配装置对上述取出的原始文档的字串和上述最佳路径中的字串进行选配。如图 16（6）所示，用标记符号（*表示丢字，[ ]表示错别字）将在上述选配时发现的不同的地方作上标记。然后，由输出部 700 输出上述最佳路径中的字串和作了标记后的字串。

以上根据实施例说明了本发明，但本发明并不限定于上述实施例，在不变更其主要意思的范围内，当然可以适当方式变形后来实施。即，例如词典中的读音符号可以直接用 2 字节的压缩符号表示。另外，也可以将字音词典和破音字典合并起来使用。

如上所述，如果采用本发明的汉语文档自动校正方法及其装置，则能解决现有的问题，能获得如下效果。

（1）能有效地对汉语文档进行错误检测和修正等。从由小学课本构成的词库取出实验数据一万字。然后，人为地造成有错误的文档后，记录该有错误的位置。如果采用本发明，则错误检测率及修正率至少能达到 87%以上。

（2）不需要准备语言模型和特殊的知识数据库。进而在知识数据库的收集和维护方面能节省大量的工时和经费。

（3）能应用于汉语输入法或文字识别装置中的后处理。例如，欲输入“流血事件是可怕的。”这一字串时，直至“流血事件”是正确的变换，但如果输入“是可怕的。”时，就会产生“流血是见识可怕的。”这样的错误变换。如果采用本发明，则在如上所述那样输入后继词的话，

可解决正确地变换过的前面的词被误变换的问题。

由以上所述可知，本发明的实用性非常强。

说 明 书 附 图

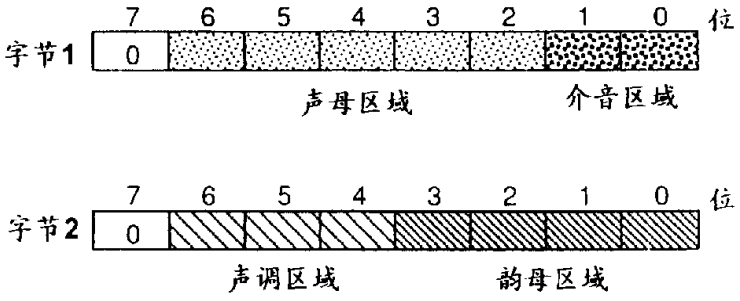

7 6 5 4 3 2 1 0 位
字节1
0
声母区域
介音区域
7 6 5 4 3 2 1 0 位
字节2
0
声调区域
韵母区域

图 1

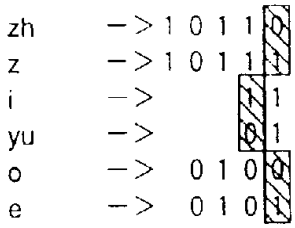

zh ->1011
z ->1011
i ->1
yu ->1
o ->0100
e ->0101

图 2

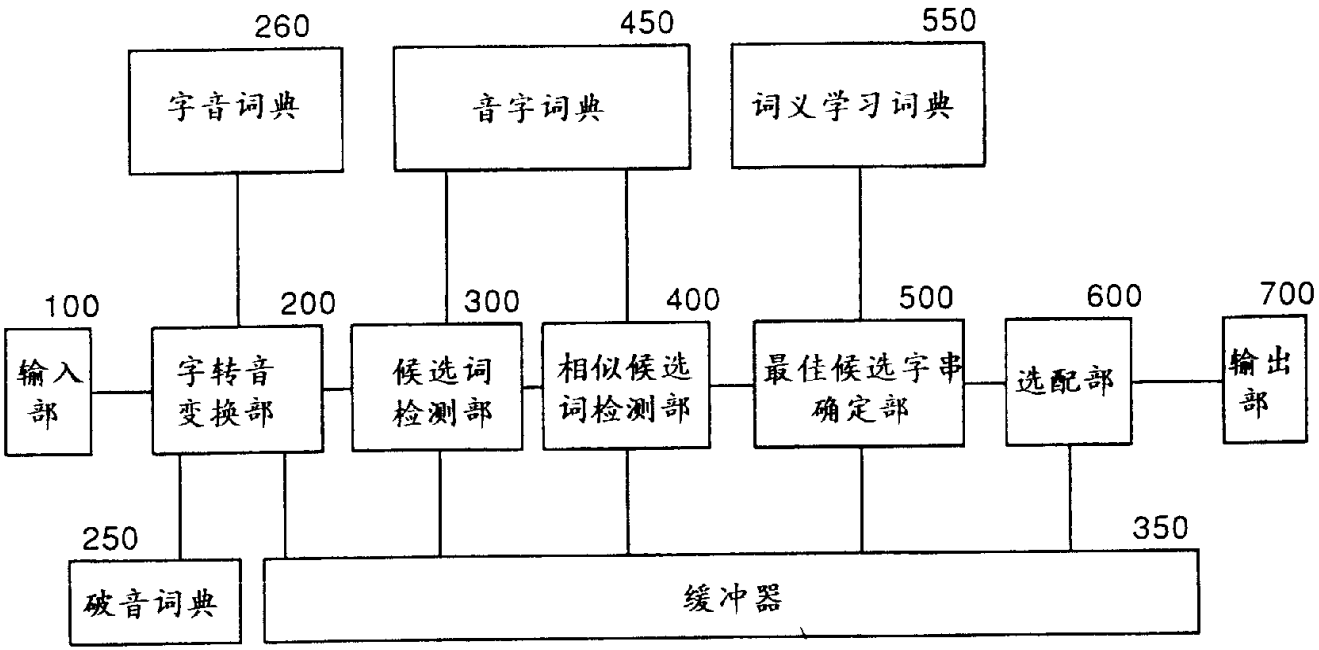

260
字音词典
450
音字词典
550
词义学习词典
100
输入部
200
字转音变换部
300
候选词检测部
400
相似候选词检测部
500
最佳候选字串确定部
600
选配部
700
输出部
250
破音词典
350
缓冲器

图3

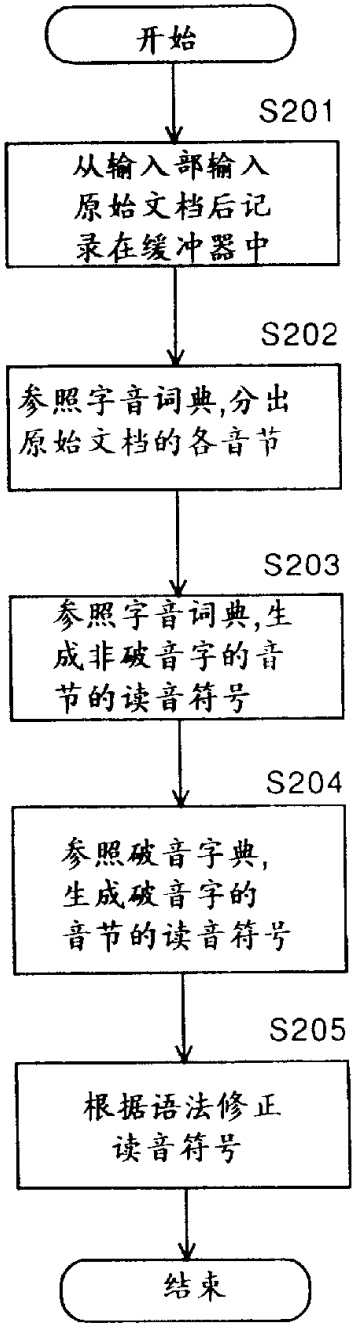

开始
S201
从输入部输入
原始文档后记
录在缓冲器中
S202
参照字音词典,分出
原始文档的各音节
S203
参照字音词典,生
成非破音字的音
节的读音符号
S204
参照破音字典,
生成破音字的
音节的读音符号
S205
根据语法修正
读音符号
结束

图 4

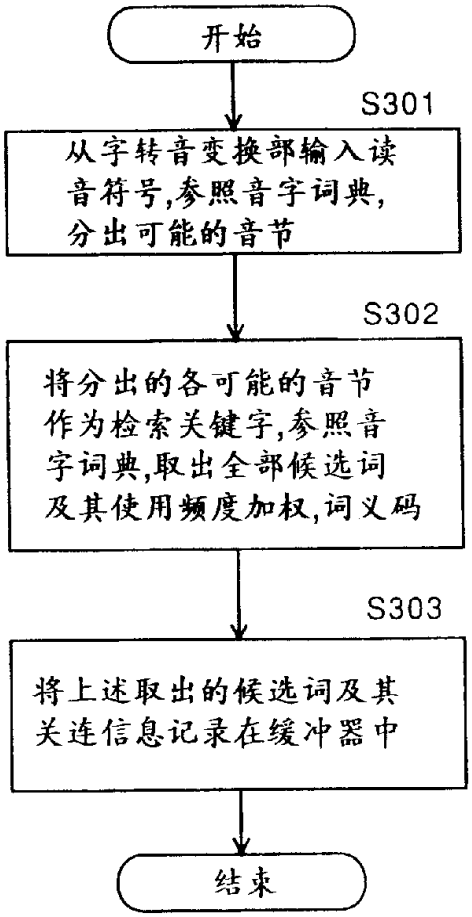

开始
S301
从字转音变换部输入读音符号,参照音字词典,分出可能的音节
S302
将分出的各可能的音节作为检索关键字,参照音字词典,取出全部候选词及其使用频度加权,词义码
S303
将上述取出的候选词及其关连信息记录在缓冲器中
结束

图 5

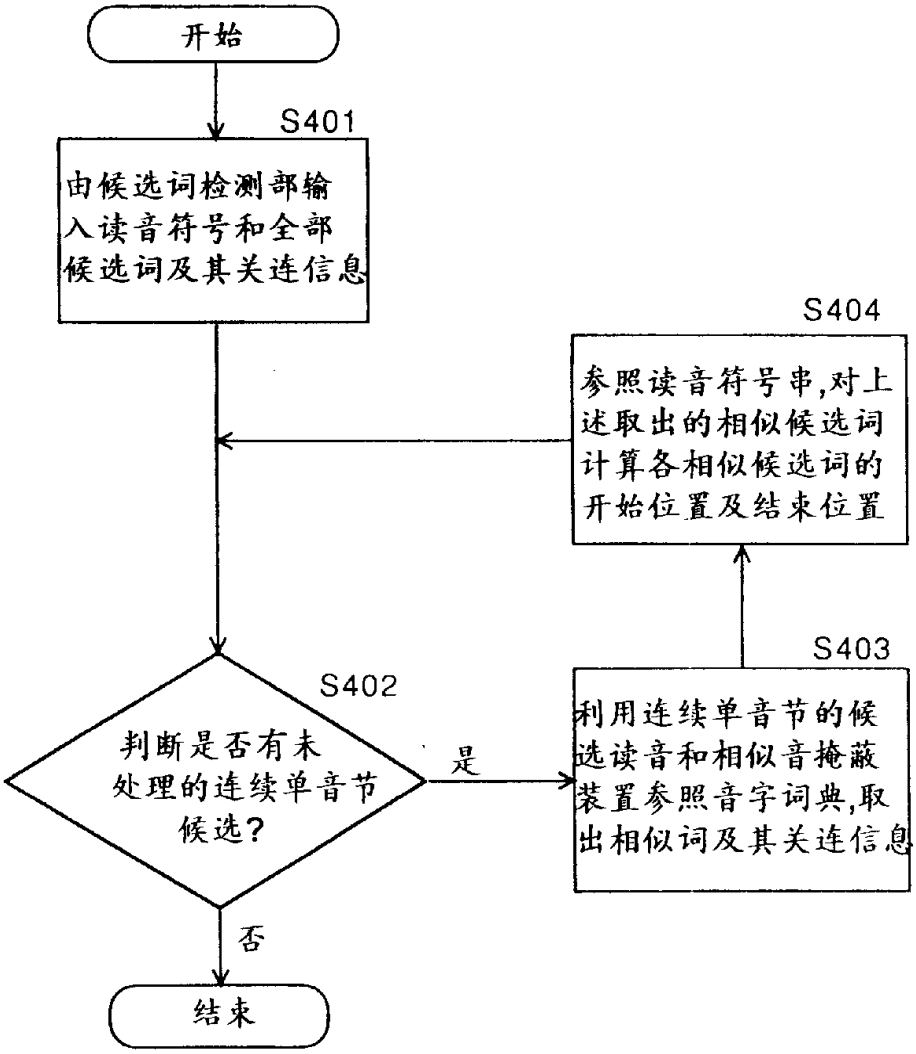

开始
S401
由候选词检测部输入读音符号和全部候选词及其关连信息
S404
参照读音符号串,对上述取出的相似候选词计算各相似候选词的开始位置及结束位置
S403
利用连续单音节的候选读音和相似音掩蔽装置参照音字词典,取出相似词及其关连信息
S402
判断是否有未处理的连续单音节候选?
是
否
结束

图 6

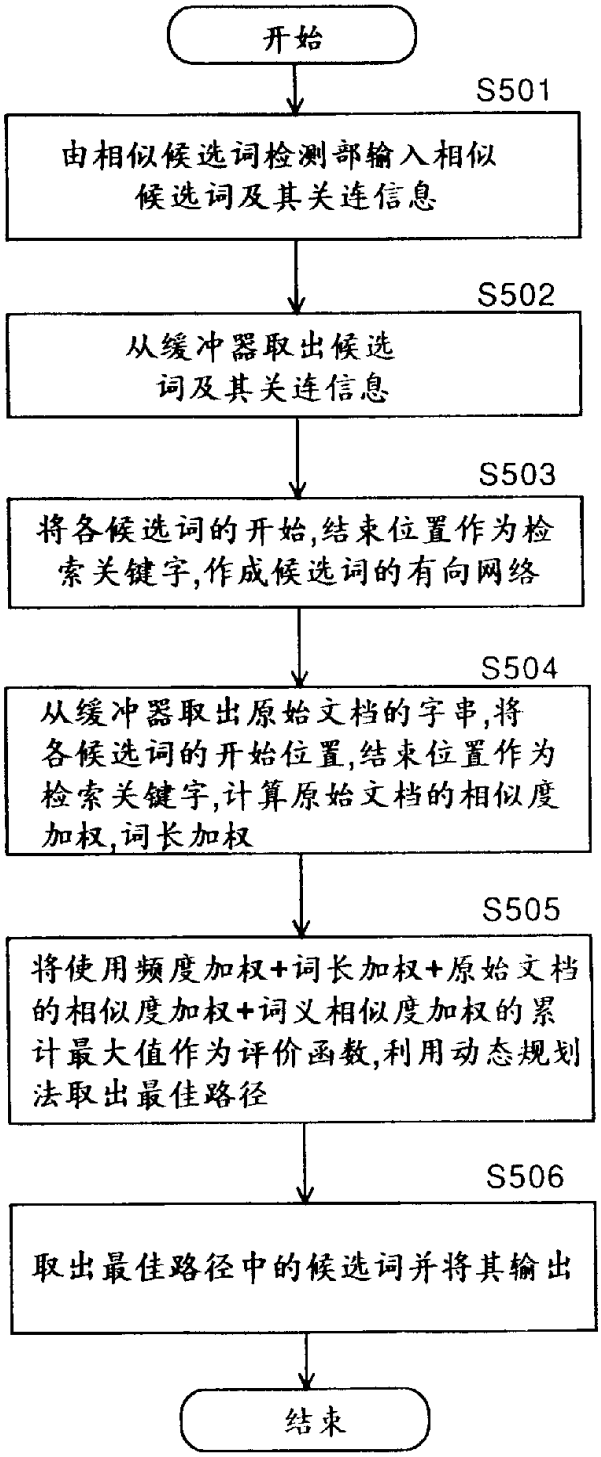

开始
S501
由相似候选词检测部输入相似候选词及其关连信息
S502
从缓冲器取出候选词及其关连信息
S503
将各候选词的开始,结束位置作为检索关键字,作成候选词的有向网络
S504
从缓冲器取出原始文档的字串,将各候选词的开始位置,结束位置作为检索关键字,计算原始文档的相似度加权,词长加权
S505
将使用频度加权+词长加权+原始文档的相似度加权+词义相似度加权的累计最大值作为评价函数,利用动态规划法取出最佳路径
S506
取出最佳路径中的候选词并将其输出
结束

图 7

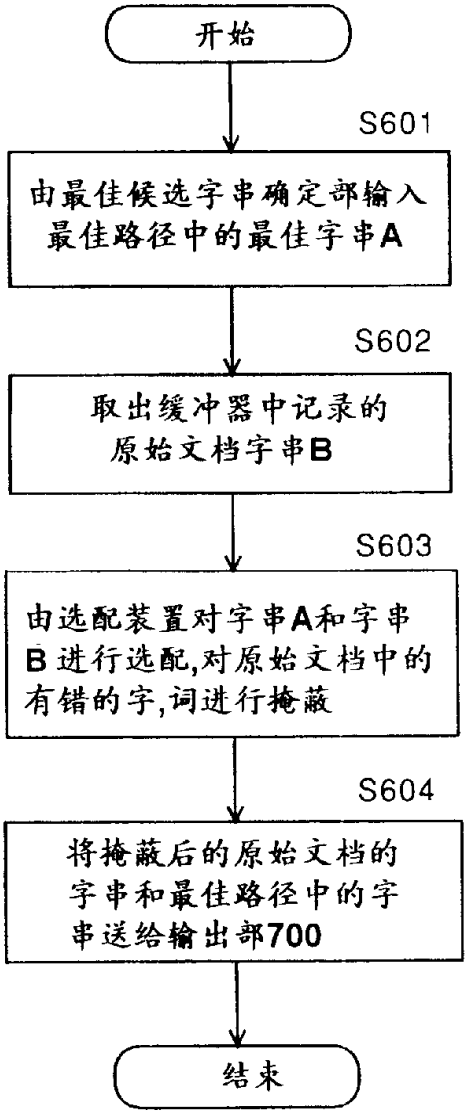

开始
S601
由最佳候选字串确定部输入最佳路径中的最佳字串A
S602
取出缓冲器中记录的原始文档字串B
S603
由选配装置对字串A和字串B进行选配,对原始文档中的有错的字,词进行掩蔽
S604
将掩蔽后的原始文档的字串和最佳路径中的字串送给输出部700
结束

图 8

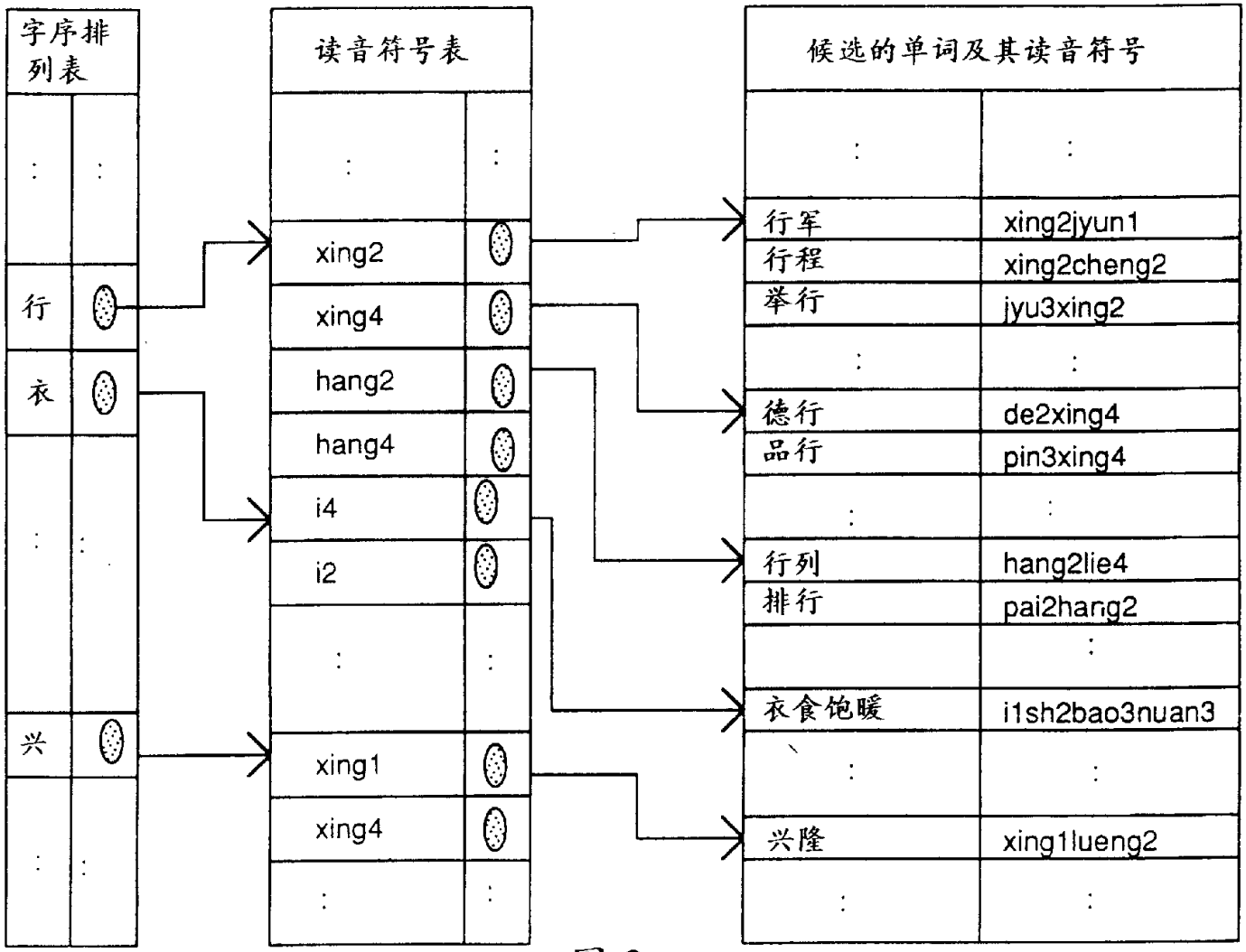

字序排列表
读音符号表
候选的单词及其读音符号
行
衣
兴
xing2
xing4
hang2
hang4
i4
i2
xing1
xing4
行军
xing2jyun1
行程
xing2cheng2
举行
jyu3xing2
德行
de2xing4
品行
pin3xing4
行列
hang2lie4
排行
pai2hang2
衣食饱暖
i1sh2bao3nuan3
兴隆
xing1lueng2

图 9

文字符号	错误的读音符号	其它可能的读音符号
· · 多	· · duo1	· · duo2
· · 建	· · jian4	
键	jian4	
兴	xing1	xing4
语	yu3	yu4
行	xing2	xing4,hang2,hang4
· · 库	· · ku4	
料	liao4	
· · 资 · ·	· · z1 · ·	

图 10

读音符号	同音异义词及其关连使用频度加权和词义码
.	.
.	.
duo1	多,0.99,126a, 朵,0.01,828e
jian4	键,0.01,9810,建,0.18,.203b, 见,0.12,3310,件,0.21,828e,...
yu3	与,0.35,833a, 语0.15,8300,雨,0.25,0230,羽 0.07,0.65a,...
z1	姿,0.1,1100,资,0.25,747b, 兹0.02,324a,...
liao4	瞭,0.15,3310,料,0.32,748a,...
ku4	酷,0.5,933b,库,0.2,940d,...
xi4	细,0.27,117d, 系,0.08,1840,...
tueng3	统,0.18,2670,筒,0.07,953b,...
. .	. .
duo1jian4	多件,1,828e
z1liao4	资料,1,8050
z1liao3ku4	资料库,1,940d
jian4yu3	键语,1,822b
xi4tueng3	系统,1,1310
. .	. .

图 11

后继词的词义码	可能的前一词的词义码
.	.
.	.
016b	828h
1310	369,200,227a,940d,....
1360	464a
.	.
.	.
822b	126a
828e	126a,120b
.	.
.	.
940d	822b,369

图 12

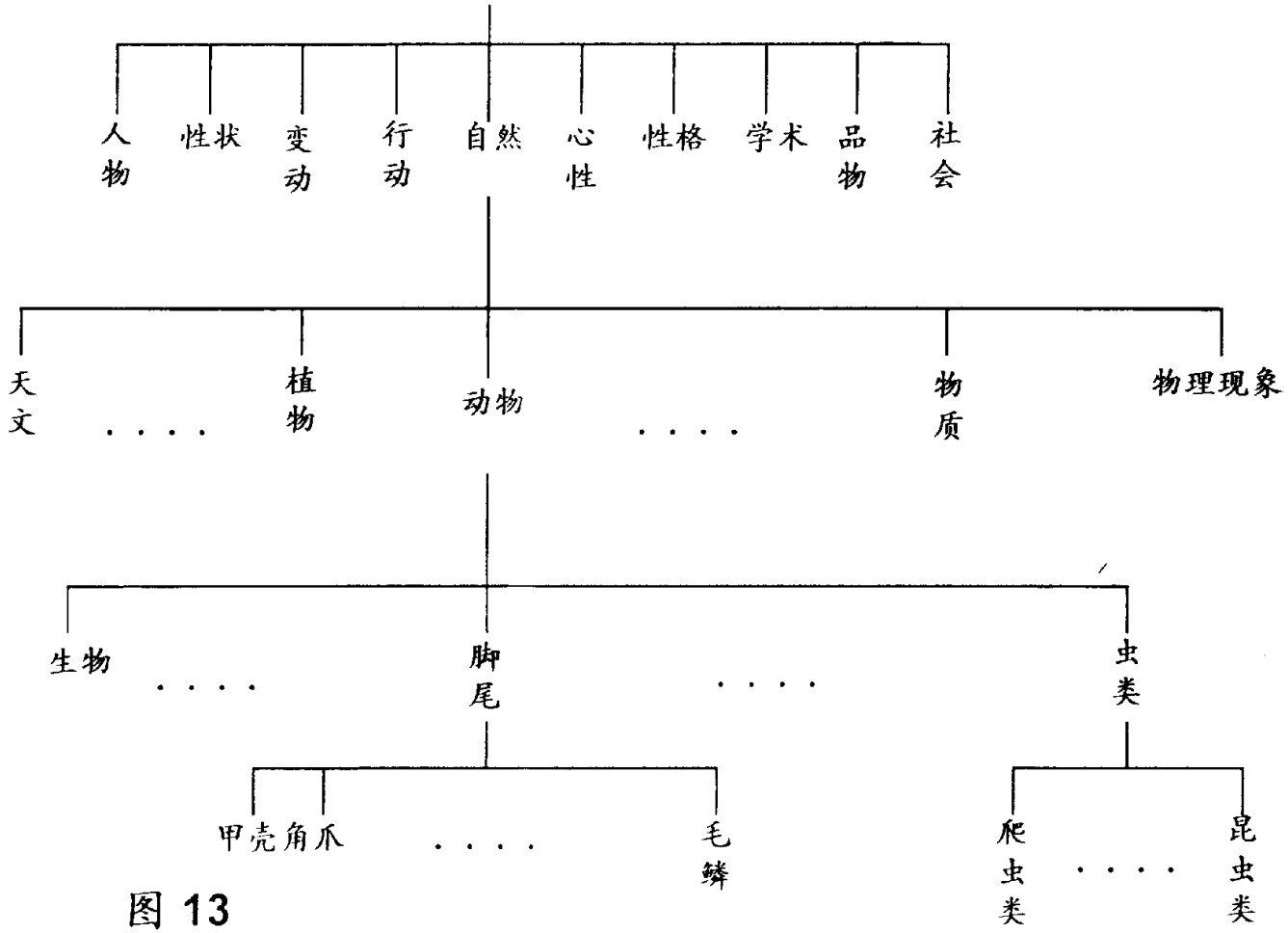
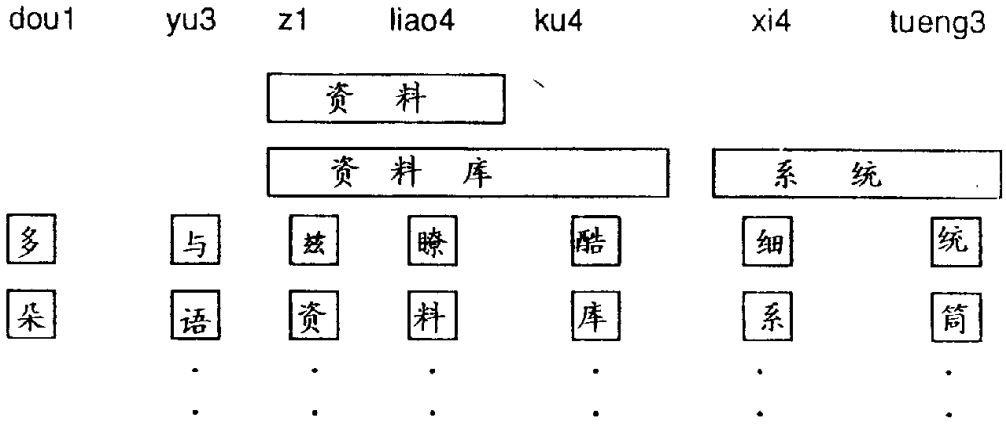

人物
性状
变动
行动
自然
心性
性格
学术
品物
社会
天文
植物
动物
物质
物理现象
生物
脚尾
虫类
甲壳
角
爪
毛鳞
爬虫类
昆虫类

图 13

（1）原始文档的字串:多词资料库系统

（2）字转音变换部的处理结果:**duo1yu3z1liao4ku4xi4tueng3**

（3）候选词检测部中的候选词:



（4）相似候选词检测部中的候选词

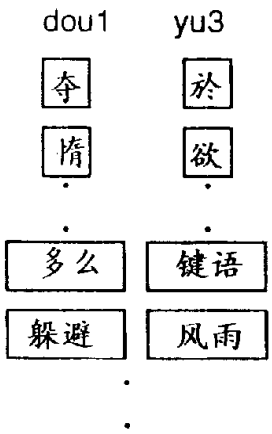


图 14

（5）最佳候选词确定部

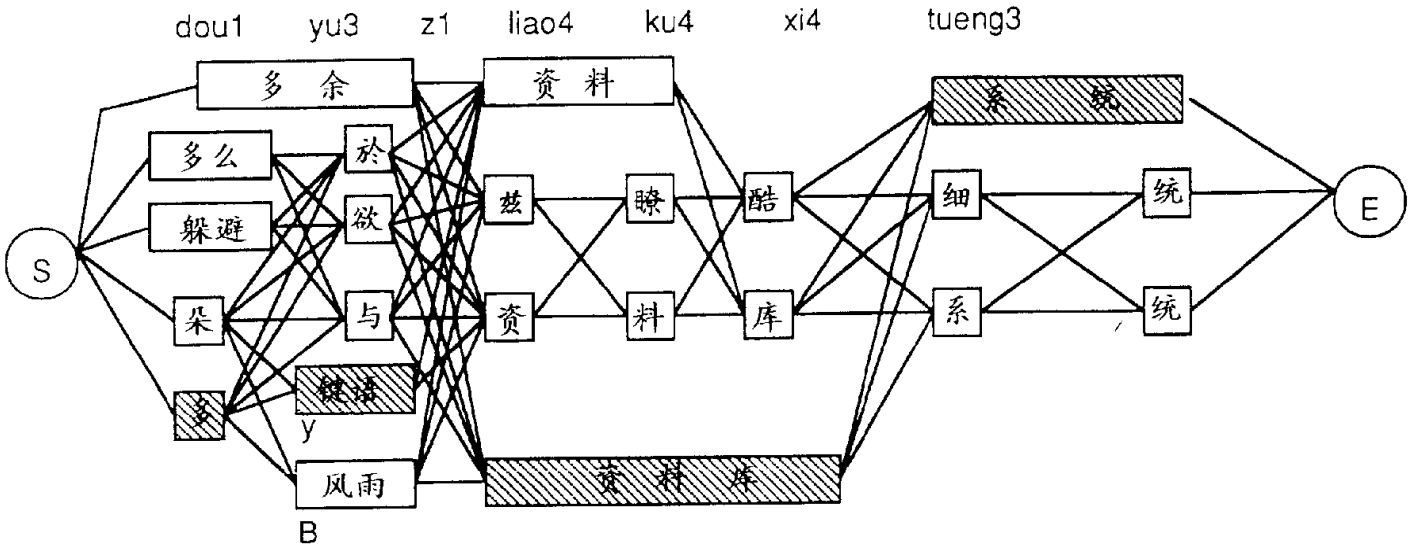


图 15

（6）选配部

原始文档的字串：多词资料库系统

最佳路径中的字串：多键词资料库系统

掩蔽后的原始文档的字串：多＊词资料库［系］统

＊：丢字、［］：错别字

（7）输出部

最佳路径中的字串：多键词资料库系统

掩蔽后的原始文档的字串：多＊词资料库［系］统

＊：丢字、［］：错别字

图 16

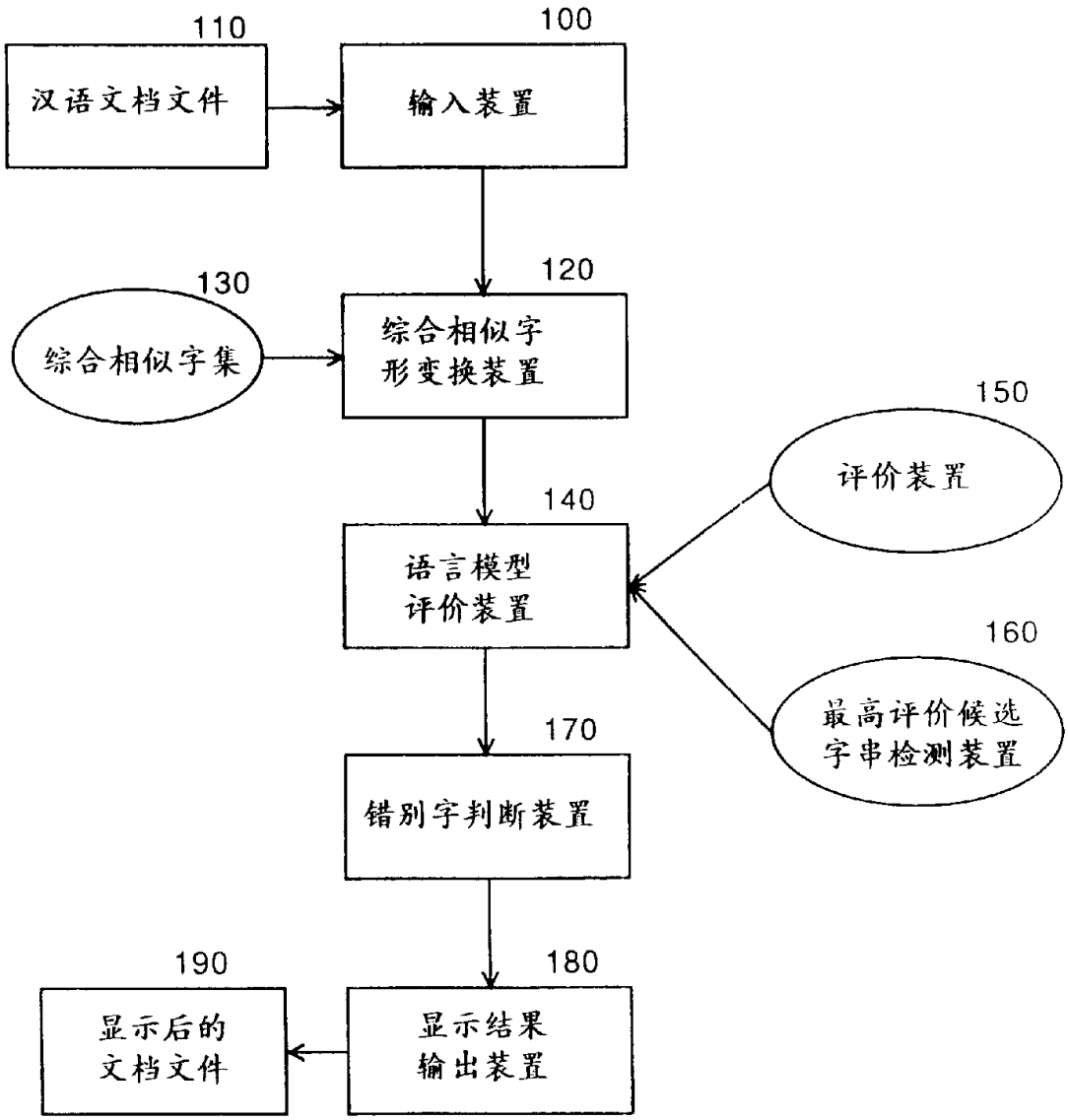

110
汉语文档文件
100
输入装置
120
130
综合相似字集
综合相似字
形变换装置
150
评价装置
140
语言模型
评价装置
160
最高评价候选
字串检测装置
170
错别字判断装置
190
显示后的
文档文件
180
显示结果
输出装置

图 17