[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G06F 17/28 (2006.01)
G06F 3/023 (2006.01)
H04M 1/23 (2006.01)

[12] 发明专利申请公开说明书

[21] 申请号 200510006708.9

[43] 公开日 2006 年 8 月 9 日

[11] 公开号 CN 1815467A

[22] 申请日 2005.1.31
[21] 申请号 200510006708.9
[71] 申请人 日电（中国）有限公司
地址 100738 北京市东城区东长安街 1 号东方广场 E3 座 1201 室
[72] 发明人 许荔秦 薛敏宇

[74] 专利代理机构 中科专利商标代理有限责任公司
代理人 罗松梅

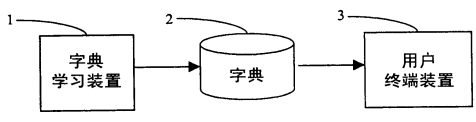
权利要求书 7 页 说明书 17 页 附图 18 页

[54] 发明名称

字典学习方法以及使用该方法的装置，输入方法以及使用该方法的用户终端装置

[57] 摘要

本发明公开了一种字典学习方法，所述方法包括步骤：从未标注的语料中学习词典和统计语言模型；将所述词典，统计语言模型以及辅助词编码信息整合为小尺寸的字典。 此外，还公开了一种在用户终端装置上使用的输入方法以及一种用户终端装置，其中所述终端装置上装载有添加了词性信息和词性双元模型的字典。 因此，通过用户终端装置可以给出句子预测和词预测，同时通过利用字典索引中的 Patricia 树索引搜索字典从而加速输入。



1．一种字典学习方法，所述方法包括步骤：

从未标注的语料中学习词典和统计语言模型；

将所述词典，统计语言模型以及辅助词编码信息整合为字典。

2． 如权利要求 1 所述的字典学习方法，所述方法还包括步骤：

从词性已标注的语料中获得所述词典中每个词的词性信息和词性双元模型；

将所述词性信息以及词性双元模型添加到所述字典中。

3． 如权利要求 1 或 2 所述的字典学习方法，其中辅助词编码信息包括汉字编码信息。

4． 如权利要求 1 或 2 所述的字典学习方法，其中辅助词编码信息包括非汉字编码信息。

5． 如权利要求 3 所述的字典学习方法，其中汉字编码信息至少包括拼音编码信息和笔画编码信息之一。

6． 如权利要求 1 或 2 所述的字典学习方法，其中从未标注的语料中学习词典和统计语言模型的步骤包括下列步骤：

a)将未标注的语料分割为词序列；

b)利用所述词序列创建统计语言模型，其中统计语言模型包括词单元模型和词三元模型；

c)计算困惑度，并判断是否是第一次计算困惑度或者困惑度降低的数值大于第一阈值；

d)在 c)的结果是肯定的情况下根据词三元模型将语料重新分割为词序列并执行步骤 b)；

e)在 c)的结果是否定的情况下根据统计语言模型优化词典，从而添加新词并删除不重要的词；

f)更新词单元模型，删除无效的词三元模型并执行步骤 a)，直到词典不再变化。

7. 如权利要求 6 所述的字典学习方法，其中步骤 a)根据下列等式对未标注的语料进行分割：

$$\hat{S}\left\{w_1 w_2 \cdots w_{n_{\hat{s}}}\right\} = \arg\max_{s} P\left(S\left\{w_1 w_2 \cdots w_{n_s}\right\}\right),$$

其中 $S\{w_1 w_2 \cdots w_{n_s}\}$ 表示词序列 $w_1 w_2 \cdots w_{n_s}$，$P\left(S\left\{w_1 w_2 \cdots w_{n_s}\right\}\right)$ 表示所述词序列的似然概率。最优的词序列是 $\hat{S}\left\{w_1 w_2 \cdots w_{n_{\hat{s}}}\right\}$。

8. 如权利要求 7 所述的字典学习方法，其中步骤 d)包括根据词典利用最大匹配对语料进行重新分割。

9. 如权利要求 6 所述的字典学习方法，其中步骤 a)包括根据词典利用最大匹配对语料进行分割。

10. 如权利要求 9 所述的字典学习方法，其中步骤 d)包括根据词典利用最大匹配对语料进行重新分割。

11. 如权利要求 6 所述的字典学习方法，其中步骤 e)包括步骤：

e1)根据第一出现计数阈值过滤出所有的三元词条和双元词条，从而形成新候选词列表；

e2)根据互信息阈值从新候选词列表中过滤出所有的候选词作为第一候选词；

e3)针对在新候选词列表中所有的第一候选词计算相对熵，并按照相对熵降序顺序对第一候选词进行排序；

e4)根据第二出现计数阈值过滤出所述词典中的所有词，从而形成删除候选词列表；

e5)将删除候选词列表中的每个词分割为一序列所述词典中的其他词，作为第二候选词；

e6)计算删除候选词列表中的所有第二候选词的相对熵，并按照相对熵升序顺序对第二候选词进行排序；

e7)确定应该添加的第一候选词的数量以及删除的第二候选词的数量，并更新所述词典。

12. 如权利要求 11 所述的字典学习方法，其中步骤 e2)根据下列等式计算所有的候选词的互信息：

$$MI(w_1, w_2 \cdots w_n) = \frac{f(w_1, w_2 \cdots w_n)\ 30}{\sum_{i=1}^{n} f(w_i) - f(w_1, w_2 \cdots w_n)}$$

其中 $(w_1,w_2\cdots w_n)$ 是词序列， $f(w_1,w_2\cdots w_n)$ 表示词序列$(w_1,w_2\cdots w_n)$的出现频率,n 等于 2 或 3。

13．一种字典学习装置，其中所述装置包括：

用于学习字典的字典学习处理模块；

存储有未标注的语料的存储单元；

用于控制所述装置的各部分的控制单元；

其中，所述字典学习处理模块包括

词典与统计语言模型学习单元，用于从未标注的语料中学习词典和统计语言模型；

字典整合单元，用于将所述词典，统计语言模型以及辅助词编码信息整合为字典。

14．如权利要求 13 所述的字典学习装置，其中所述存储单元还存储有词性已标注的语料，以及字典学习处理模块还包括：

词性学习单元，用于从词性已标注的语料中获得所述词典中每个词的词性信息和词性双元模型；

以及字典整合单元将所述词性信息以及词性双元模型添加到字典中。

15．如权利要求 13 或 14 所述的字典学习装置，其中词典与统计语言模型学习单元通过执行下列处理从未标注的语料学习词典和统计语言模型：

将未标注的语料分割为词序列；

利用所述词序列创建统计语言模型，其中统计语言模型包括词单元模型和词三元模型；

通过词三元模型将所述语料重复分割为词序列，并利用词序列创建统计语言模型，直到不是第一次计算困惑度以及困惑度降低的数值小于第一阈值；

根据统计语言模型优化词典，从而添加新词并删除不重要的词；以及

更新词单元模型，删除无效的词三元模型并将未标注的语料分割为词序列，直到词典不再变化。

16．如权利要求 15 所述的字典学习装置，其中词典与统计语言模型学

习单元通过执行下列处理优化词典：

根据第一出现计数阈值过滤出所有的三元词条和双元词条，从而形成新候选词列表；

根据互信息阈值将从新候选词列表中过滤出所有的候选词作为第一候选词；

针对在新候选词列表中的所有的第一候选词计算相对熵，并按照相对熵降序顺序对第一候选词进行排序；

根据第二出现计数阈值过滤出所述词典中的所有词，从而形成删除候选词列表；

将删除候选词列表中的每个词分割为一序列所述词典中的其他词，作为第二候选词；

针对删除候选词列表中的所有第二候选词计算相对熵，并按照相对熵升序顺序对第二候选词进行排序；

确定应该添加的第一候选词的数量以及删除的第二候选词的数量，并更新所述词典。

17. 如权利要求 13 所述的字典学习装置，其中辅助词编码信息包括汉字编码信息。

18. 如权利要求 13 所述的字典学习装置，其中辅助词编码信息包括非汉字编码信息。

19. 如权利要求 17 所述的字典学习装置，其中汉字编码信息至少包括拼音编码信息和笔画编码信息之一。

20. 一种用于处理用户输入的输入方法，其中所述方法包括：

接收步骤，用于接收用户输入；

解译步骤，用于将用户输入解译为编码信息或用户动作，其中基于字典预先获得字典中的每个词的编码信息；

用户输入预测与调整步骤，用于在接收到编码信息或用户动作时，根据字典中的统计语言模型和词性双元模型利用字典索引中的 Patricia 树索引给出句子与词预测，并根据用户动作调整句子和词预测；

显示步骤，用于显示句子和词预测的结果。

21. 如权利要求 20 所述的用于处理用户输入的输入方法，其中接收

步骤接收汉字输入。

22. 如权利要求 20 所述的用于处理用户输入的输入方法，其中接收步骤接收非汉字输入。

23. 如权利要求 21 所述的用于处理用户输入的输入方法，其中所述汉字输入包括拼音输入，笔画输入以及笔迹输入之一。

24. 如权利要求 20 所述的用于处理用户输入的输入方法，其中用户输入预测与调整步骤包括下列步骤：

a)接收解译的编码信息或用户动作；

b)如果接收到的是用户动作则修改预测结果并执行步骤 h)；

c)根据编码信息从所有的当前 Patricia 树节点搜索所有可能的 Patricia 树索引的新 Patricia 树节点；

d)如果不存在任何新 Patricia 树节点，则忽略所述编码信息并重置所有的搜索结果以及执行步骤 a)；

e)如果存在新 Patricia 树节点，则将新 Patricia 树节点设置为当前的 Patricia 树节点；

f)从当前的 Patricia 树节点搜索所有的可能词并给出句子预测；

g)根据句子预测结果确定当前词，并给出词预测，所述词预测包括候选词列表和预测候选词列表；以及

h)输出预测结果以显示并返回执行步骤 a)。

25. 如权利要求 24 所述的用于处理用户输入的输入方法，其中步骤 f)根据下列等式确定最可能的词序列作为预测句子从而给出句子预测：

$$\hat{S}\left(w_1 w_2 \cdots w_{n_{\hat{S}}}\right) = \arg\max_{S} \sum_{t_1 \in POS_{w_1}, t_2 \in POS_{w_2}, \cdots} P\left(S\left(w_1 o_{t_1} w_2 o_{t_2} \cdots w_{n_s} o_{t_{n_s}}\right) \mid I\right),$$

$$P(S) = P\left(O_{t_1}\right)\frac{P(w_1)P\left(O_{t_1} \mid w_1\right)}{P\left(O_{t_1}\right)} P\left(O_{t_2} \mid O_{t_1}\right)\frac{P(w_2)P\left(O_{t_2} \mid w_2\right)}{P\left(O_{t_2}\right)} \cdots P\left(O_{t_{n_s}} \mid O_{t_{n_s-1}}\right)\frac{P\left(w_{n_s}\right)P\left(O_{t_{n_s}} \mid w_{n_s}\right)}{P\left(O_{t_{n_s}}\right)},$$

其中

POS_{w_1} 是词 w_1 所具有的所有词性的集合；

O_{t_n} 是词 w_n 的一个词性；

$P(O_{i_1})$ 和 $P(O_{i_2}|O_{i_1})$分别是词性单元和词性双元;

$P(w_1)$ 是词单元；以及

$P(O_{i_1}|w_1)$ 是一个词对应词性的概率。

26. 一种用于处理用户输入的用户终端装置，其中所述装置包括：

用户输入终端，用于接收用户输入；

存储单元，用于存储字典和包括 Patricia 树索引的字典索引；

输入处理单元，用于根据用户输入给出句子和词预测；以及

显示器，用于显示句子和词预测的结果；

其中，输入处理单元包括

输入编码解译器，用于将用户输入解译为编码信息或用户动作，其中基于字典预先获得字典中的每个词的编码信息；

用户输入预测与调整模块，用于在接收到编码信息或用户动作时，根据字典中的统计语言模型和词性双元模型利用字典索引中的 Patricia 树索引给出句子和词预测，并根据用户动作调整句子和词预测。

27. 如权利要求 26 所述的用于处理用户输入的用户终端装置，其中输入处理单元还包括字典加索引模块，用于给出字典中每个词条的编码信息，根据编码信息和词单元对所有词条进行排序，构建 Patricia 树索引并将其添加到字典索引中。

28. 如权利要求 26 和 27 所述的用于处理用户输入的用户终端装置，其中用户输入预测与调整模块通过执行下列处理给出句子和词预测并调整句子和词预测：

接收解译的编码信息或用户动作；

如果接收到的是用户动作则修改预测结果并将结果输出显示；

如果接收到的是编码信息，则根据编码信息从所有的当前 Patricia 树节点搜索所有可能的 Patricia 树索引的新 Patricia 树节点；

如果不存在任何新 Patricia 树节点，则忽略所述编码信息并重置所有的搜索结果，然后重复执行接收解译的编码信息或用户动作；

如果存在新 Patricia 树节点，则将新 Patricia 树节点设置为当前的 Patricia 树节点；

从当前的 Patricia 树节点搜索所有的可能词并给出句子预测；

根据句子预测结果确定当前词，并给出词预测，所述词预测包括候选词列表和预测候选词列表；以及

输出预测结果以显示。

29. 如权利要求 26 所述的用于处理用户输入的用户终端装置，其中用户输入终端用于汉字输入。

30. 如权利要求 26 所述的用于处理用户输入的用户终端装置，其中用户输入终端用于非汉字输入。

31. 如权利要求 29 所述的用于处理用户输入的用户终端装置，其中用户输入终端可以是数字键盘，其中每个数字按键代表拼音编码。

32. 如权利要求 29 所述的用于处理用户输入的用户终端装置，其中用户输入终端可以是数字键盘，其中每个数字按键代表笔画编码。

33. 如权利要求 29 所述的用于处理用户输入的用户终端装置，其中用户输入终端可以是触摸板。

字典学习方法以及使用该方法的装置，输入方法以及使用该方法的用户终端装置

技术领域

本发明涉及一种自然语言处理，更具体地，涉及一种字典学习方法以及使用该字典学习方法的装置，输入方法以及使用该输入方法的用户终端装置。

背景技术

随着计算机、PDA 以及移动电话在中国的广泛应用，可以看出这些装置的一个重要特征在于能够使用户实现中文输入。在中国目前的移动终端市场，几乎每一个移动电话都提供利用数字键盘的输入方法。当前最广泛使用的输入方法为 T9 以及 iTap。利用这种输入方法，用户可以使用十按键数字键盘输入汉字的拼音或笔画。附图 8A-8B 示出用于拼音和笔画输入的示例键盘。该输入方法根据用户敲击的按键顺序给出汉字预测。当用户输入一个汉字的拼音时，用户不需要按照最常规的输入方法点击按键三到四次输入每个正确的字母。用户仅需要根据该汉字的拼音点击一系列按键，则输入方法就会在一个候选列表中预测出正确的拼音和正确的汉字。例如，用户想利用拼音“jin”输入“今”，他不需要通过敲击 “5” (代表 “jkl”) 1 次来输入 “j”，敲击 “4” (代表 “ghi”) 3 次以及敲击 “6” (代表“mno”) 2 次，然而，他仅需敲击 “546”则输入方法将给出预测拼音“jin” 以及对应的预测候选汉字 “进今金...”. 图 9A 示出利用最传统的输入方法输入汉字“今”的 T9 的输入序列。

对于当前的移动终端来说，用户必须逐字地输入汉字。虽然一些输入方法宣称可以根据用户输入给出预测结果，但实际上，这些输入方

法是逐字地给出预测的。对于每个汉字，用户需要点击按键若干次，并至少进行一次拼写核对。鉴于此，本发明的发明人提供一种可以给出句子级以及词级的预测的系统。

如上所述，目前T9和iTap是移动终端上最为广泛使用的输入方法。然而，这些输入方法的速度不能够令大多数的用户满意。需要多次点击以及多次交互，即使仅输入单个汉字。

存在上述问题的主要原因在于应用中文输入方法的当前大部分数字键盘仅仅是基于汉字的（US 20030027601）。这是因为在汉字中，在词之间并不存在清晰的界限。此外，对词也没有明确的定义。因此，这些输入方法选择将单个汉字看作是与其英文相对应的“词”。然而，这将不可避免地导致依据单个汉字的数字序列的大量的冗余汉字，速度也因此明显的降低。此外，由于仅能根据单个汉字获得预测，所以基于汉字的输入方法在很大程度上限制了词预测的效果。也就是说，当前移动终端中所采用的输入方法仅能够将用户输入的数字序列转换为汉字候选列表。用户必须从候选列表中选出正确的汉字。用户不能够连续地输入一个词或一个句子。

例如，用户想输入词“今天”。首先，用户使用数字键盘输入“546”，其表示汉字“今”的拼音“jin”。 然后，向用户显示候选列表“进今金...”。其次，用户必须从该列表中选出正确的汉字“今”. 然后，向用户显示可以跟随在汉字“今”之后的候选列表“天日年...”。用户必须从该列表中选出正确的汉字“天”。图9B示出输入汉字词“今天”的T9的输入序列。

在PC平台中，存在基于PC键盘的多种高级快速输入方法，诸如微软拼音，紫光拼音以及智能狂拼等。其中的一些方法可以给出句子级的预测，所有的上述方法可以给出词级的预测。但是对于这些可以给出句子级预测的方法来说，字典的尺寸太大。例如，微软拼音输入的字典大小为20～70 MB, 智能狂拼所需要的存储空间达到100MB。它们都采用统计语言模型（SLM）技术来形成可以进行句子预测的基于词的SLM（典型地是词双元模型或词三元模型）。然而这种SLM使用了预定的词典并在字典中存储了大量的词双元词条和词三元词条，字

典的尺寸将会不可避免地太大，从而不能够安装在移动终端上。此外，在移动终端平台上的预测速度也非常慢。

另一个不利之处在于大多数的输入方法没有词典或仅包括预定的词典。因此，不能够连续地输入在语言中频繁使用的多个重要的词和短语，如“今天下午”。

发明内容

因此，考虑到上述问题提出本发明，以及本发明的目的是提供一种字典(dictionary)学习方法和利用该字典学习方法的装置。此外，本发明也提供一种输入方法以及一种使用该输入方法的用户终端装置。该装置从语料中学习字典。学习的字典包括优化的词典(lexicon)，该词典包括多个从语料中学习的重要的词以及短语。然而，在该字典应用到随后描述的输入方法中时，它还包括词性信息以及词性双元模型。用户终端装置使用 Patricia 树（一种树状的数据结构）索引搜索字典。所述装置接收用户输入并基于字典搜索的结果给出句子和词预测，所述词预测包括当前候选词列表和预测候选词列表。向用户显示预测结果。所以，用户通过连续地输入与词或句子相对应的数字序列可以输入词或句子。从而用户不需要针对每个汉字输入数字序列并从候选词列表中选出正确的汉字。因此输入速度得到了很大改善。

根据本发明的第一方面，提供了一种字典学习方法，所述方法包括步骤：从未标注的语料中学习词典和统计语言模型；将所述词典，统计语言模型以及辅助词编码信息整合为字典。

根据本发明的第二方面，所述字典学习方法还包括步骤：从词性已标注的语料中获得所述词典中每个词的词性信息和词性双元模型；将所述词性信息以及词性双元模型添加到字典中。

根据本发明的第三方面，提供了一种字典学习装置，其中所述装置包括：用于学习字典的字典学习处理模块；存储有未标注的语料的存储单元；用于控制所述装置的各部分的控制单元；其中，所述字典学习处理模块包括词典与统计语言模型学习单元，用于从未标注的语料中学习词典和统计语言模型；字典整合单元，用于将所述词典，统

计语言模型以及辅助词编码信息整合为字典。

根据本发明的第四方面，其中所述字典学习装置的存储单元还存储有词性已标注的语料，以及字典学习处理模块还包括：词性学习单元，用于从词性已标注的语料中获得所述词典中每个词的词性信息和词性双元模型；以及字典整合单元将所述词性信息以及词性双元模型添加到字典中。

根据本发明的第五方面，提供了一种用于处理用户输入的输入方法，其中所述方法包括：接收步骤，用于接收用户输入；解译步骤，用于将用户输入解译为编码信息或用户动作，其中基于字典预先获得字典中的每个词的编码信息；用户输入预测与调整步骤，用于在接收到编码信息和用户动作时，根据字典中的统计语言模型和词性双元模型利用词典索引中的 Patricia 树给出句子与词预测，并根据用户动作调整句子和词预测；显示步骤，用于显示句子和词预测的结果。

根据本发明的第六方面，提供了一种用于处理用户输入的用户终端装置，其中所述装置包括：用户输入终端，用于接收用户输入；存储单元，用于存储字典和包括 Patricia 树索引的字典索引；输入处理单元，用于根据用户输入给出句子和词预测；以及显示器，用于显示句子和词预测的结果；其中，输入处理单元包括：输入编码解译器，用于将用户输入解译为编码信息或用户动作，其中基于字典预先获得字典中的每个词的编码信息；用户输入预测与调整模块，用于在接收到编码信息和用户动作时，根据字典中的统计语言模型和词性双元模型利用词典索引中的 Patricia 树索引给出句子和词预测，并根据用户动作调整句子和词预测。

根据本发明，通过利用具有小尺寸的字典可以给出句子级预测和词级预测。其中所述的字典通过本发明第四方面的字典学习装置的学习处理而获得。所述字典学习装置从语料中提取大量的重要信息，并将其以特定内容和特定结构的形式保持，从而可以以非常小的尺寸进行存储。与移动电话上的常规输入方法不同，本发明的基本输入单元是“词”。这里所述的“词”也包括从语料中学习的“短语”。根据所述字典的内容和结构，输入方法可以给出句子级和词级的预测。因此，

与常规输入方法例如 T9 和 iTap 相比较，输入速度增加。

虽然与基于 PC 的输入方法比较，例如与微软拼音比较，其中微软拼音可以给出句子和词预测但是使用了尺寸较大的字典从而存储预定的词典以及对应的大量的词双元词条或词三元词条，本发明学习的字典仅在优化的词典中存储提取出的重要的语言信息以及对应词单元的字典。因此，在字典中的所有信息对于语言处理来说都是必要的信息，这些信息仅需要较小的存储成本。

如下详细地描述了本发明的优点：

1.可以学习包括优化的词典的字典。所述优化的字典包括从语料中学习的多个重要的词和短语。

2.所述学习的字典包括优化的字典以及一些词性信息。帮助给出句子和词预测的所述字典的尺寸非常小，从而能够在移动电话上使用。

3.利用 Patricia 树索引为字典编索引。所述 Patricia 树索引的使用有助于快速地搜索词。因此，可以很容易地、快速地获得句子和词预测。

附图说明

通过对参考附图的下列优选实施例的具体描述，本发明的上述特征和优点将会变得非常明显，其中：

图 1 示出描述了本发明的字典学习装置和用户终端装置之间的关系的示意图；

图 2A 示出了由字典学习装置学习的字典的示意结构的示例；

图 2B 示出了由字典学习装置学习的字典的示意结构的另一示例；

图 3 示出了根据本发明的字典学习装置的方框图；

图 4A 示出了字典学习装置的字典学习处理模块的示例的详细的方框图；

图 4B 示出了字典学习装置的字典学习处理模块的另一示例的详细的方框图；

图 5 是一流程图，用于解释根据本发明的字典学习处理模块中的词典与统计语言模型学习单元执行的学习词典以及统计语言模型的过

程；

图 6 是根据本发明的词典优化的流程图；

图 7 示出根据本发明第一实施例的用户终端装置的方框图；

图 8A-8D 示出用户终端装置的四个常规键盘的示意框图；

图 9A 示出利用最常规的输入方法输入汉字“今”时 T9 的输入序列；

图 9B 示出利用最常规的输入方法输入汉字“今天”时 T9 的输入序列；

图 10 示出在本发明的用户终端装置的输入处理单元的不同部分之间的连接关系的方框图；

图 11 示出本发明的用户终端装置的显示器的用户界面的示例；

图 12 示出由本发明用户终端装置的字典加索引模块执行的构建 Patricia 树索引的流程图；

图 13 示出本发明排序结果和 Patricia 树索引的示例；

图 14 示出由本发明用户终端装置的用户输入预测与调整模块执行的用户输入预测以及调整的过程的流程图；

图 15 示出用户终端装置的输入序列的示例；

图 16 示出根据本发明第二实施例的用户终端装置的方框图。

具体实施方式

下面将参考附图 1 描述示出了本发明的字典学习装置和用户终端装置之间的关系的示意图。字典学习装置 1 学习计算机可读字典 2。用户终端装置 3 使用字典 2 帮助用户输入文本。字典学习装置 1 和用户终端装置 3 相互独立。字典学习装置 1 训练的字典 2 还可以用于其它的应用。字典学习装置 1 使用特定的字典学习方法以及特定的字典结构，以构建向用户提供快速输入的小尺寸的字典。

图 2A 示出了由字典学习装置学习的字典的示意结构的示例。在该示例中，部分 2 包括多个词条 (部分 21)。所述的词条不仅用于“词”（例如，“打扮”），而且是“短语”（例如，“打扮整齐”，“打扮整齐”，“打扮整齐干净”）。所述“短语”实际上是一复合词（由一序列的词构

成)。为了避免在下述描述中的不便，术语“词”指的是传统的“词”和传统的“短语”两者。一些其它词的例子包括“今天”，“今天下午”，“今天下午八点”。部分 21 包括词干（术语“词干”和本发明中所述的词的含义相同）（部分 211），词单元（部分 212），该词所具有的若干词性（部分 213）以及这些词性的对应概率（部分 214），一些辅助词编码信息（部分 215)。部分 215 可以是拼音（汉字的发音）编码信息或笔画编码信息或者是其它的编码信息。将哪种类型的部分 215 添加到部分 21 中取决于字典的应用。在下面描述的示例中，也可以不包括部分 215。最后，部分 22，即词性双元模型也包含在该示例中。该部分也取决于应用并且可以不包括在其它的示例中。正如对本领域的普通技术人员所显而易见的是字典 2 并不局限于汉字，它可以是任何类型的字典。对于日语，字典的所有部分与汉字相同，除了辅助词编码信息（部分 215）应该是平假名编码信息而非拼音编码信息。例如，对于词“今晚”，平假名编码信息是“こんばん”。对于英语，字典的所有部分与汉字相同，除了应该省略辅助词编码信息，因为英语单词编码信息就是该单词的字符序列。对于韩语，字典的所有部分与汉字相同，除了辅助词编码信息（部分 215）应该是韩语笔画编码信息，而不是拼音编码信息。例如，对于词“휴대폰”， 韩语笔画编码信息是“ㅎㅡ. . ㄷㅣ. ㅣㅍ.ㅡㄴ”。随后将描述由图 4A 所示的示例装置如何学习该字典。

图 2B 示出了由字典学习装置学习的字典的示意结构的另一示例。与图 2A 所示的示例相比较，在该示例中的字典不包括该词的词性（部分 213)，这些词性的对应概率（部分 214）以及词性双元模型（部分 22)。该字典的使用范围可以比第一示例的范围更广。它可以用在手写、语音识别后处理、输入方法以及其它语言相关的应用中。随后将描述由图 4B 所示的示例装置如何学习该字典。

下面将参考图 3 和图 4A 来描述学习图 2 所示的字典的字典学习装置 1。如图 3 和图 4A 所示，字典学习装置 1 包括通过内部总线 103 连接的 CPU 101，附件 102，存储器 104 以及硬盘 105。存储器 104 存储操作系统 1041，字典学习处理模块 1042 以及其它的应用程序

1043。硬盘 105 存储语料 1051，字典学习文件 1052 以及其它的文件（未示出）。由字典学习装置 1 学习的字典 2 也存储在硬盘上。语料 1051 包括，例如，未标注的语料 12 和词性已标注的语料 1051。字典学习文件 1052 包括词典 11 和统计语言模型 14。字典学习处理模块 1042 包括词典与统计语言模型学习单元 15，词性学习单元以及字典整合单元 17。

由字典学习处理模块 1042 训练生成最后的字典 2。字典学习处理模块 1042 读取语料 1051 并将词典 11 以及统计语言模型 14 写在硬盘上并在硬盘上输出最终的字典 2。

词典 11 由词干的集合组成。起初，包括语言中的传统词的普通词典可以用作词典 11。词典与统计语言模型学习单元 15 将学习最终的词典和统计语言模型，同时在此过程中对词典 11 进行优化。删除词典 11 中的一些不重要的词以及添加一些重要的词和短语。未标注的语料 11 是包括大量没有分割为词序列的文本但包括多个句子的文本语料（对于英语，一个句子可以通过一些例如空格的“标记”而分割为“词”序列。但是这些“词”仅仅是传统“词”，而不是包括了在本说明书中所称的“词”的传统“短语”）。词典与统计语言模型学习单元 15 处理词典 11 以及未标注的语料 12，然后创建统计语言模型 14（初始并不存在）。统计语言模型 14 包括词三元模型 141 以及词单元模型 142。然后，词典与统计语言模型学习单元 15 使用统计语言模型 14 中的信息来优化词典 11。词典与统计语言模型学习单元 15 重复这一处理过程并创建最终的词典 11 以及最终的词单元模型 142。

词性已标注的语料 13 是利用对应词性标注词序列的语料。典型地，可以手工创建该语料，但其规模受到了限制。词性学习单元 16 扫描词性已标注的语料 13 的词序列。基于词典 11，词性 16 为词典中的每一个词统计词性信息。计数一个词的所有词性以及其对应概率（字典 2 中的部分 213）。对于词典 11 中没有在词序列中出现的词，手工地给予该词一个词性以及给出其对应的概率 1。在该过程中利用传统的双元模型计算方法给出词性双元模型（字典 2 中的部分 22）。

通过使用词单元模型 142，词三元模型 141，词典 11，以及词性

学习单元 16 给出的一些词性信息，字典整合单元整合上述的所有数据并添加一些应用程序所需的辅助词编码信息（字典 2 中的部分 215），从而创建图 2A 中所描述的最终的字典 2。

下面将参考图 3 和图 4B 描述学习字典的字典学习装置的另一示例。与图 3 和图 4B 所示的示例相比较，语料 1051 仅包括已标注的语料。字典学习处理模块 1042 不包括词性学习单元 16。因此，在该示例中并不考虑词性相关的信息。字典整合单元 17 将词三元模型 141，词单元模型 142，词典 11 以及一些应用程序所需的辅助词编码信息（字典 2 中的部分 215）整合为如图 2B 所示的最后的字典 2。

图 5 是一流程图，用于解释由词典与统计语言模型学习单元 15 执行的学习词典以及统计语言模型的过程。首先，在步骤 151 将未标注的语料 12 分割为词序列。对于该分词步骤存在多种不同的方法。第一种方法是仅根据词典使用最大匹配来分割语料 12。第二种方法是：在词单元模型 142 存在的情况下，根据词单元模型 142 利用最大似然来分割语料 12；在词单元模型 142 不存在的情况下，根据词典利用最大匹配来分割语料 12。最大似然是一种分词的标准方法，如等式（1）所示：

$$\hat{S}\left\{w_1 w_2 \cdots w_{n_{\hat{s}}}\right\} = \arg\max_{s} P\left(S\left\{w_1 w_2 \cdots w_{n_s}\right\}\right) \qquad (1)$$

在等式 (1)中，$S\{w_1 w_2 \cdots w_{n_s}\}$表示词序列 $w_1 w_2 \cdots w_{n_s}$。 $P(S\{w_1 w_2 \cdots w_{n_s}\})$表示该词序列的似然概率。优化的词序列为 $\hat{S}\left\{w_1 w_2 \cdots w_{n_{\hat{s}}}\right\}$。

在步骤 152，接收分割的词序列，以及利用常规的 SLM 创建方法创建统计语言模型 14，其中所述统计语言模型包括词三元模型 141 以及词单元模型 142。

在步骤 153，使用步骤 152 中创建的词三元模型评价在步骤 151 产生的词序列的困惑度（Perplexity）。如果是第一次计算困惑度，则

处理直接进行到步骤 154。否则，将新获得的困惑度与旧的困惑度相比较。如果新的困惑度降低的数值超过了预定的阈值，则处理进行到步骤 154；否则处理进行到步骤 155。

在步骤 154，根据新创建的词三元模型 141 利用最大似然来将语料 12 重新分割为词序列，并执行步骤 152。

在步骤 155，根据统计语言模型中的一些信息将一些新词添加到词典中并从词典中删除一些不重要的词，从而优化了词典。在下面的段落中将描述如何进行词典优化。一个新词通常是词三元模型 141 中的三元词条或双元词条的词序列组成的新词。例如，如果“今天”，“下午”和“八点“都是当前词典中的词，则双元词条“今天 下午”或者三元词条“今天 下午 八点”可能成为优化后的词典中的新词。如果这两个词都被添加了，则优化后的词典应该包括词“今天 下午”以及词“今天 下午 八点”。

在步骤 156，评价词典。如果在步骤 155 词典并没有改变（没有添加新词也没有删除不重要的词），则词典与统计语言模型学习单元 15 停止该处理。否则该处理进行到步骤 157。

在步骤 157，由于词三元模型 141 和词单元模型 142 与新创建的词典不再对应，因此词三元模型 141 和词单元模型 142 不再有效。此时根据新的词典更新词单元模型；从词三元模型得到新词的词单元出现概率；并且删除被删除的词单元词条。最后，删除词三元模型 141 并重复执行步骤 151。

图 6 示出了根据本发明的词典优化的流程图。当词典优化开始时，存在两条要执行的路径。一条是执行步骤 1551，另一条是执行步骤 1554。可以选择任何一条路径先执行。

首先，在步骤 1551，利用出现计数阈值过滤出所有的三元词条（例如“今天 下午 八点”）以及双元词条（例如“今天 下午”），例如，在语料中出现次数超过 100 的所有词条都被选择到新词候选列表中。由此创建了一个新词候选列表。在步骤 1552，通过互信息阈值过滤出所有的候选词。如下定义了互信息：

$$MI(w_1,w_2\cdots w_n)=\frac{f(w_1,w_2\cdots w_n)}{\sum_{i=1}^{n}f(w_i)-f(w_1,w_2\cdots w_n)} \qquad (2)$$

其中 $f(w_1,w_2\cdots w_n)$ 表示词序列 $(w_1,w_2\cdots w_n)$的出现频率. 这里 $(w_1,w_2\cdots w_n)$ 作为新候选词, n 等于 2 或 3。例如，对于 w_1今天, w_2下午 以及 w_3八点，候选词“今天 下午 八点”的互信息是 $MI(\text{今天下午八点})=\frac{f(\text{今天下午八点})}{f(\text{今天})+f(\text{下午})+f(\text{八点})-f(\text{今天下午八点})}$。从候选词列表中删除互信息小于阈值的所有候选词。

在步骤 1553，为新候选词列表中的每个候选词计算相对熵。如下定义了相对熵：

$$D(w_1,w_2,\cdots,w_n)=f(w_1,w_2,\cdots,w_n)\log\left[\frac{P(w_1,w_2,\cdots,w_n)}{f(w_1,w_2,\cdots,w_n)}\right] \qquad (3)$$

其中 $P(w_1,w_2,\cdots,w_n)$是当前词三元模型给出的词序列 $(w_1,w_2\cdots w_n)$的似然概率。然后在步骤 1553，按照相对熵的降序顺序排序所有的候选词。

在进行到步骤 1557 之前，必须首先处理右边的路径（步骤 1554~1556）。右边的路径是删除一些不重要的词(例如 “革命委员会”)以及一些 “伪词”。当将一词序列添加为新词时，它可能是“伪词” (例如 “今天下”)。因此，需要删除一些词典词条。

在步骤 1554，通过出现计数阈值过滤出所有的词，例如，在词典中出现次数小于 100 的所有词都被选择到删除词候选列表中。由此创建了一个包括要删除的候选词的删除候选词列表。

在步骤 1555，将删除候选词列表中的每个词分割为其它的词序列。例如，将革命委员会” 分割为 “革命”,“委员会”。 该分词方法与步骤 151 或步骤 154 所描述的分词方法类似。可以使用这两个步骤中的任何一种方法。

与步骤 1553 类似，在步骤 1556 计算每个候选词的相对熵。然后，以相对熵的升序顺序排序所有的候选词。

在步骤1557，采用策略依据两个候选词列表来确定应该添加多少新候选词以及应该删除多少候选词，所述候选词列表是：一个是有关新词的列表，另一个是有关删除词的列表。所述策略可以是一个规则或多个规则。例如，使用相对熵的阈值，或使用词典中的词的总数作为判断手段，或者使用上述这两种判断手段。最后，更新该词典。

如何进行词典优化是非常重要的。在词典优化过程中，将初始仅是一些词序列的重要的短语添加到词典中作为新词，因此，可以将在初始的词单元模型中并不存在的一些重要的语言信息提取到最终的词单元模型中。并且，从初始的词单元模型中删除一些不重要的语言信息。所以，最终的词单元模型可以保持有小尺寸而在进行语言预测时却具有更好的性能。这也是本发明能够提供一种小尺寸的词典的同时能在进行句子和词的预测时具有良好性能的重要的原因。

图7示出了根据本发明第一实施例的用户终端装置的方框图。如图7所示，由总线34连接处理器31，用户输入终端32，显示器33，RAM 35以及ROM（闪存）36并使其交互作用。输入处理单元3601中包括输入编码解译器362，字典加索引模块363，用户输入预测与调整模块364。在ROM 36上装载有输入处理单元3601，字典2，字典索引366，操作系统361以及其它的应用程序365。

图 8A-8D 示出本发明所采用的用户终端装置的四个常规键盘的示意框图。用户输入终端 32 可以是任何类型的用户输入装置。如图8A所示，一个示例的用户输入终端32是数字键盘，其中每个数字按键代表拼音编码。按键321是数字“4”，代表拼音字符“g”或“h”或“i”。按键322是功能键，用户可以使用这种按键进行一些动作。例如，点击该按键若干次从而从候选列表中选出正确的候选词。所述的示例的用户输入终端也可以应用于英文输入。因此每个数字按键代表若干字母表字符。用户输入终端32的另一个例子是图8B所示的数字键盘，其中每个数字按键代表若干笔画编码。在图8B中，按键321是数字“4”，代表笔画“、”。用户输入终端32的第三个例子是日语输入所采用的数字键盘。在该例中，每个数字按键代表若干平假名。在图 8C中，按键321是数字“4”，代表平假名“た”或“ち”或“つ”或“て”或

“と”。 用户输入终端 32 的第四个例子是用于韩文输入的数字键盘。在该例中，每个数字键盘代表若干韩语笔画。在图 8D 中，按键 321 是数字“4“，代表韩语“ㄱ” 或 “ㅋ” 或“ㄲ”。 用户输入终端 32 的第五个例子是可以记录笔迹的触摸板。通过某些触摸屏的笔可以记录用户的一些动作。

图 10 示出了图 7 所示的用户终端装置的输入处理单元中的不同部分之间的连接关系的方框图。在用户输入预测与调整模块 364 工作之前，字典加索引模块 363 读取字典 2 并将字典索引 366 加到 ROM 36 中。字典索引 366 是基于对应词编码信息的字典 2 中的所有词条的索引。对于第一个示例的用户输入终端 32，词的编码信息是数字序列。例如，词“今天”的拼音是“jintian”, 所以其编码信息是 “5468426”。对于第二个示例的用户输入终端 32, 词的编码信息是数字序列。例如，词 “今天” 的笔画是 “丿、 一——丿、 ”, 因此其编码信息为“34451134”. 对于第三个示例的用户输入终端 32, 词的编码信息也是数字序列。例如, 词 “今晚” 的平假名是 “こんばん”, 因此编码信息是 “205#0”。对于第四个示例的用户输入终端 32，词的编码信息是数字序列。例如，词 “휴 대 폰” 的韩语笔画是“ㅎㅡ. . ㄷㅣ. ㅣㅍ.ㅡㄴ”, 因此编码信息为“832261217235”。对于第五个示例的用户输入终端 32, 词的编码信息 是 Unicode(统一的字符编码标准)序列。例如， 词 “今天” 的 Unicode 是 “(4ECA) (5929)”, 所以编码信息为“(4ECA) (5929)”。

用户输入终端 32 接收用户输入并将其通过总线 34 发送到输入编码解译器 362。输入编码解译器 362 将用户输入解译为编码信息或用户动作，并将其传送到用户输入预测与调整模块 364。该编码信息可以是确定的或者是随机的。对于第一个示例的用户输入终端 32，输入编码解译器 362 将每个按键点击解译为确定的数字代码（“0” ~ ”9”），代表几个可能的拼音字符(“a”~”z”)。对于第二个示例的用户输入终端 32，输入编码解译器 362 将每个按键点击解译为确定数字代码 (“0” ~ ”9”) ，代表笔画字符(“—”~” 一”)。对于第三个示例的用户输入终端 32，输入编码解译器 362 将每个按键点击解译为确定数字代码(“0” ~ ”9”以及 “#”) ，代表几个可能的平假名。 对于第四个示例的用户输

入终端 32，输入编码解译器 362 将每个按键点击解译为确定数字代码（“0”～”9”），代表几个可能的韩语笔画。对于第五个示例的用户输入终端 32，输入编码解译器 362 将每个笔迹解译为随机变量，其表示若干可能的 Unicode 以及对应概率。（输入编码解译器 362 可以是手写识别引擎，其将笔迹识别为一组候选汉字以及对应的概率）。

用户输入预测与调整模块 364 接收由输入编码解译器 362 发送的已解译的编码信息或用户动作。基于词典 2 和词典索引 366，产生用户输入结果并将其通过总线 34 发送到显示器 33。显示器 33 向用户显示输入方法产生的结果以及与该输入方法相关的其它信息。图 11 示出了用户终端装置的显示器 33 的用户界面。

该显示器所显示的用户界面包括输入状态信息区域 331 以及输入结果区域 332。在区域 331，显示了用户输入 3311 和输入方法状态 3312。区域 3311 指示已经由用户输入的当前数字序列。区域 3312 指示当前输入方法是拼音的数字键盘输入方法。在区域 332，显示了用户输入预测与调整模块 364 给出的结果。句子预测 3321 是由用户输入预测与调整模块 364 根据输入的数字序列 3311 的阴影部分（当前词部分）给出的所有当前候选词的列表。在该列表中的所有候选词具有相同的词编码信息，即，数字序列“24832”。当前的预测候选词 3323 是有关所有预测的当前候选词的列表，预测候选词 3323 由用户输入预测与调整模块 364 根据输入的数字序列 3311 的阴影部分（当前的词部分）给出。在该列表中所有候选词的词编码信息的头五个数字具有相同的数字序列“24832”。(出发点”248323426”, 厨房”2483234”, 出访”2483234”)。可以改变该显示器 33 的用户界面的布局以及可以去除或改变每个组成部分。

图 12 示出了由字典加索引模块 363 执行的构建 Patricia 树索引的流程图。在步骤 3631，字典加索引模块 363 读取字典 2。根据特定的用户输入终端，给出每个词的编码信息。然后，在步骤 3632，首先根据词条的编码信息对词条进行排序。如果两个词条的编码信息是相同的，则利用词单元进行排序。根据排序结果，构建该字典的 Patricia 树索引。Patricia 树索引可以存储大量的记录并提供对记录的快速连续

的搜索。最后，将 Patricia 树索引写入字典索引中。

图 13 示出了本发明排序结果和 Patricia 树索引的示例。通过上述的 Patricia 树索引使用字典索引 366，用户输入预测与调整模块 364 在接收到新的用户输入动作时执行快速的词搜索。例如，首先给出“2”，用户输入预测与调整模块 364 一步就可以搜索到节点“2”，并将该节点记录在存储器中。在下一步，当输入“3”时，用户输入预测与调整模块 364 仅一步就从节点“2”搜索到节点“23”。在每个节点中，可以很容易地获得用于计算对应的候选词和预测候选词的信息。

图 14 示出由本发明用户终端装置 1 的用户输入预测与调整模块 364 执行的用户输入预测以及调整的过程的流程图。在步骤 3641，接收来自输入编码解译器 362 的用户输入信息并判断该信息是用户动作还是编码信息。如果是用户动作信息，则将执行步骤 3648。否则将执行步骤 3642。

在步骤 3642，使用用户输入编码信息，并根据该编码信息沿字典索引 366 的 Patricia 树索引向前递推一步。这意味着用户输入预测与调整模块 364 存储了当前 Patricia 树节点的列表。当添加新的编码信息时，使用列表中的节点作为起始点，步骤 3642 顺着 Patricia 树索引向前递推一步以搜索新的 Patricia 树节点。如果新的编码信息为添加的初始编码信息，则步骤 3642 从 Patricia 树的根节点开始。也就是说，对于图 12 中的示例 Patricia 树，如果“2”为输入的初始编码信息，步骤 3642 从根节点开始检索 Patricia 树中的新节点“2”。然后，将“2”和根节点设置为当前的 Patricia 树节点。如果“3”为输入的第二编码信息，在步骤 3642，从当前节点“2”检索新节点“23”以及从当前节点中的根节点检索新节点“3”。最后，将节点“23”，节点“3”以及根节点设置为当前节点。

在步骤 3643，如果没有搜索到新的节点，则处理进行到步骤 3644。这意味着该编码信息无效。否则，处理进行到步骤 3645。

在步骤 3644，忽略该编码信息并重置所有的结果和状态为未加入此信息前的值。然后，处理返回到步骤 3641 等待下一用户输入信息。

在步骤 3645，接收新的 Patricia 树节点，并将其设置为当前的

Patricia 树节点。每个当前节点表示所有输入编码信息 的可能的当前词的集合。然后在该步骤进行句子预测，从而确定最有可能的词序列。最有可能的词序列是最终的句子预测。例如，分别将“2”和“3”添加为第一和第二用户输入编码信息。当前节点是“23”，“3”以及根节点。具有编码信息“23”的词是仅具有一个词的词序列。这也是一种可能的句子（“测”是可能的句子）。具有编码信息“3”的词可以在具有编码信息“2”的词之后并形成两个词序列“2”－“3”。这是另一种可能的句子（“阿 恶”为可能的句子，“啊 恶”也是可能的句子）。如何确定最可能的句子可以表述为：给出编码序列 I，找出与 I 相对应的最可能的词序列 $S(w_1 w_2 \cdots w_{n_s})$。根据等式（4）可以解决这一问题：

$$\hat{S}\left(w_1 w_2 \cdots w_{n_{\hat{S}}}\right) = \arg\max_{s} \sum_{i_1 \in POS_{w_1}, i_2 \in POS_{w_2}, \cdots} P\left(S\left(w_1 o_{i_1} w_2 o_{i_2} \cdots w_{n_s} o_{i_{n_s}}\right) \mid I\right) \tag{4}$$

POS_{w_1} 是词 w_1 所具有的所有词性的集合。O_{i_n} 是词 w_n 的词性之一。

由于需要使 $P(S)$ 最大化，可以根据等式 (5)求出 $P(S)$:

$$P(S) = P(O_{i_1}) \frac{P(w_1)P(O_{i_1} \mid w_1)}{P(O_{i_1})} P(O_{i_2} \mid O_{i_1}) \frac{P(w_2)P(O_{i_2} \mid w_2)}{P(O_{i_2})} \cdots P(O_{i_{n_s}} \mid O_{i_{n_s-1}}) \frac{P(w_{n_s})P(O_{i_{n_s}} \mid w_{n_s})}{P(O_{i_{n_s}})} \tag{5}$$

$P(O_{i_1})$ 和 $P(O_{i_2} \mid O_{i_1})$ 分别是词性单元和词性双元。它们包含在词性双元模型中 (在图 2 示出的词典 2 的部分 22)。$P(w_1)$ 是词单元 (字典 2 中的部分 212)。$P(O_{i_1} \mid w_1)$ 是一个词对应词性的概率(字典 2 的部分 214)。

在步骤 3646，确定在句子预测中的当前词。在步骤 3646，根据

该词的 Patricia 树节点，推出当前候选词和预测的当前候选词。例如，假设句子预测是“阿 恶”，当前词是“恶”。则针对当前词的 Patricia 树节点是节点“3”。因此，当前候选词列表仅包括一个词“恶”，而预测的当前候选词列表中没有词。

最后，在步骤 3647 输出要显示的结果，处理返回到 3641 等待下一个用户输入信息。

如果用户输入信息是用户动作，则步骤 3648 根据结果采取一些对应的调整。例如，如果用户从当前候选词列表中选择第二个词，则应该将句子预测中的当前词改变为根据所选择的词的新的当前词。例如，如果用户根据该句子预测结果点击“F2”（意指 OK），则将如图 11 所示的句子预测 3321 发送到当前的用户应用程序，并清除区域 332 中的数字序列 331 以及所有的结果。

图 15 示出使用图 8A 所示的键盘的用户终端装置 3 的示例输入序列。在该图中，用户通过第一示例的用户输入终端 32 使用拼音输入汉字“今天下午”。

图 16 示出根据本发明第二实施例的用户终端装置的方框图。该实施例示出两部分：用户终端装置和计算机。而图 7 所示的第一实施例仅包括一个移动终端。这两个实施例之间的区别在于：第二实施例的用户终端装置采用了计算机中的字典加索引模块 366。字典加索引模块 366 处理字典并将字典索引 366 输出到计算机的硬盘上。而将字典 2 和字典索引 366 装载在用户终端装置的 ROM（Flash）中。可以通过用户输入终端装置提供商所提供的工具进行装载处理。然后，用户输入预测与调整模块 364 可以像第一实施例中的用户终端装置那样工作。

从上述可以看出，虽然已经详细的描述了示例性的实施例，本领域的普通技术人员将会明白可能会有各种修改，添加以及替换，而不偏离附后的权利要求书所要求的本发明的保护范围以及本发明的精髓。

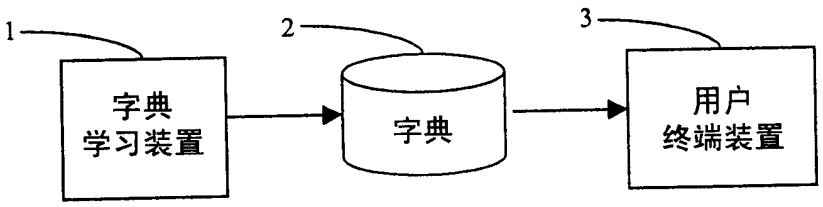


图 1

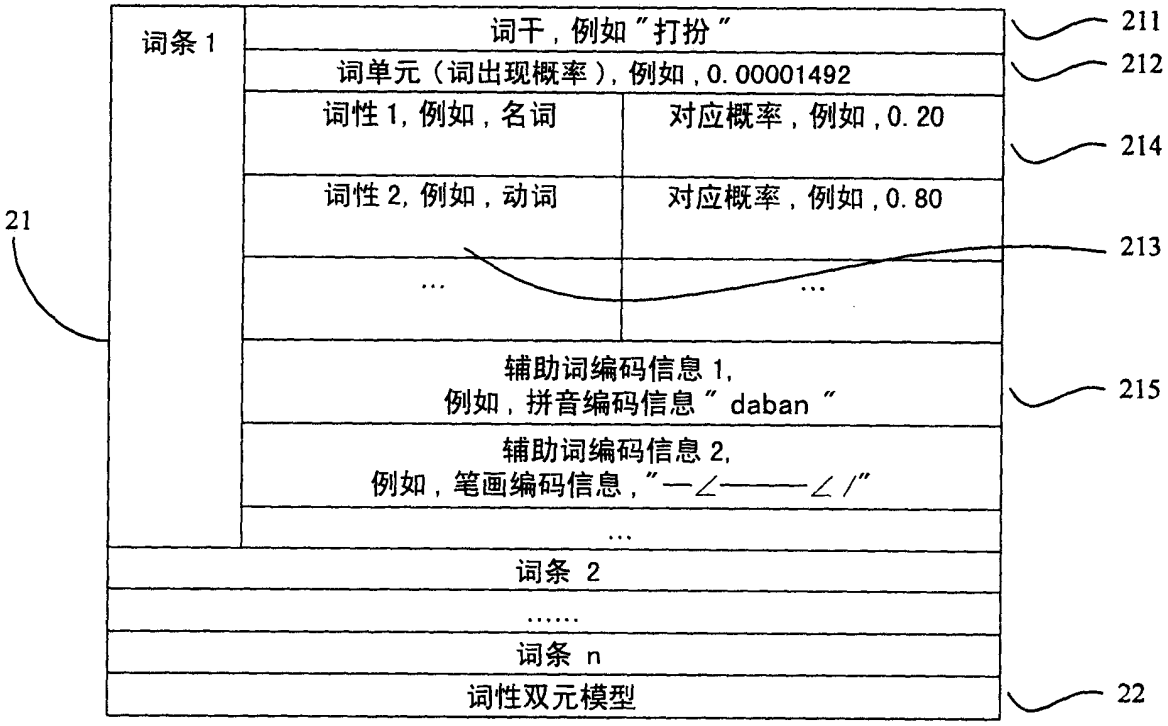

词条 1
词干，例如〝打扮〝
词单元（词出现概率），例如，0.00001492
词性 1，例如，名词
对应概率，例如，0.20
词性 2，例如，动词
对应概率，例如，0.80
…
…
辅助词编码信息 1，
例如，拼音编码信息〝 daban 〝
辅助词编码信息 2，
例如，笔画编码信息，〝一ㄥ———ㄥ丿〝
…
词条 2
……
词条 n
词性双元模型
21
211
212
214
213
215
22

图 2A

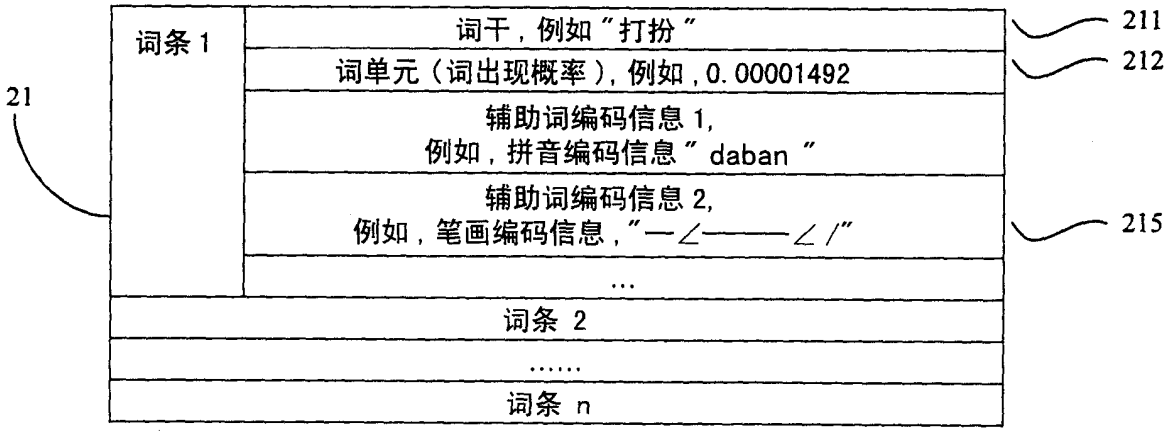

词条 1
词干，例如〝打扮〝
词单元（词出现概率），例如，0.00001492
辅助词编码信息 1，
例如，拼音编码信息〝 daban 〝
辅助词编码信息 2，
例如，笔画编码信息，〝一ㄥ———ㄥ丿〝
…
词条 2
……
词条 n
21
211
212
215

图 2B

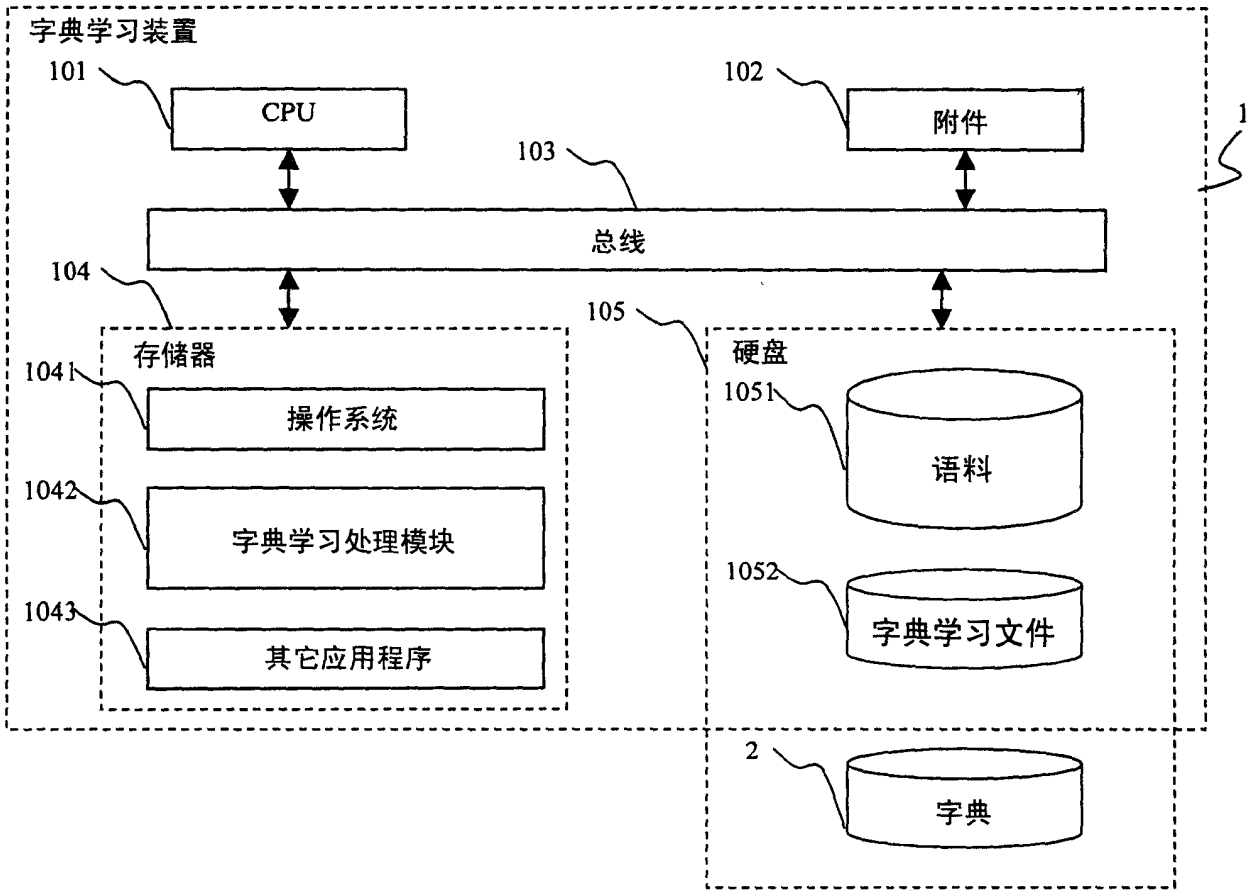

字典学习装置
101
CPU
102
附件
1
103
总线
104
105
存储器
硬盘
1041
操作系统
1051
语料
1042
字典学习处理模块
1052
字典学习文件
1043
其它应用程序
2
字典

图 3

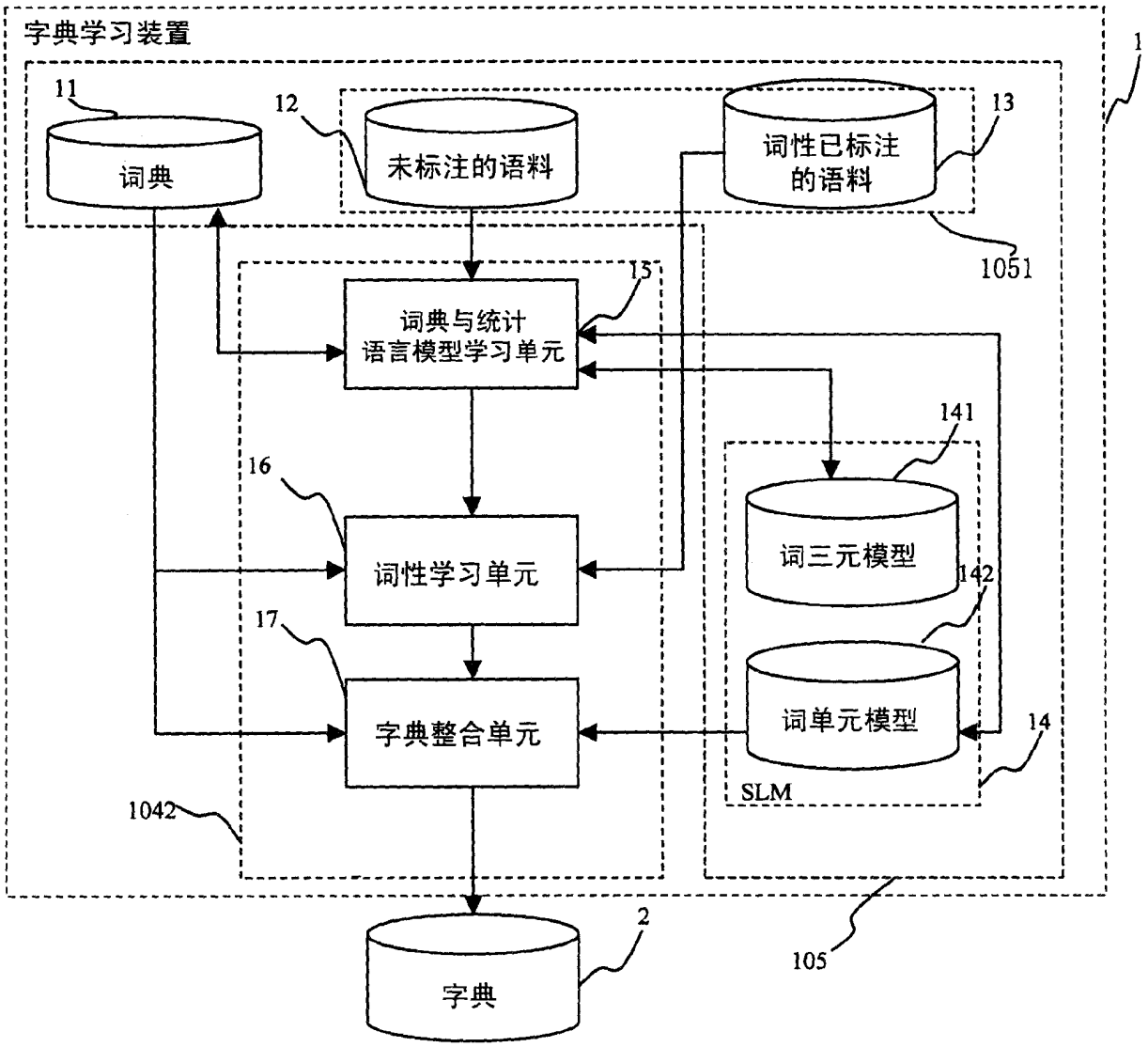

字典学习装置
1
11
词典
12
未标注的语料
13
词性已标注
的语料
1051
15
词典与统计
语言模型学习单元
141
16
词三元模型
142
词性学习单元
17
词单元模型
14
字典整合单元
SLM
1042
105
2
字典

图 4A

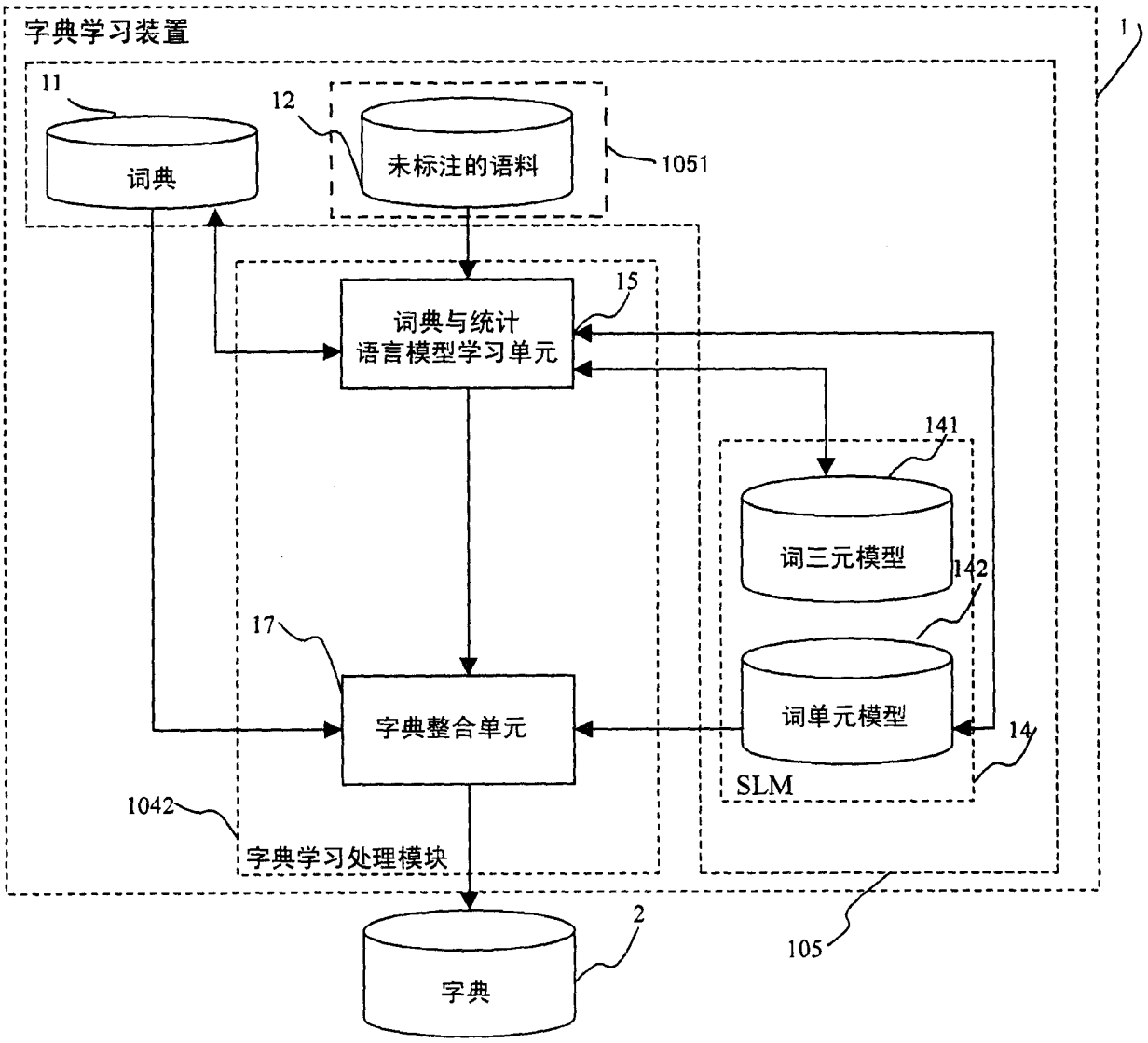

字典学习装置
1
11
词典
12
未标注的语料
1051
15
词典与统计
语言模型学习单元
141
词三元模型
142
词单元模型
14
SLM
17
字典整合单元
1042
字典学习处理模块
105
2
字典

图 4B

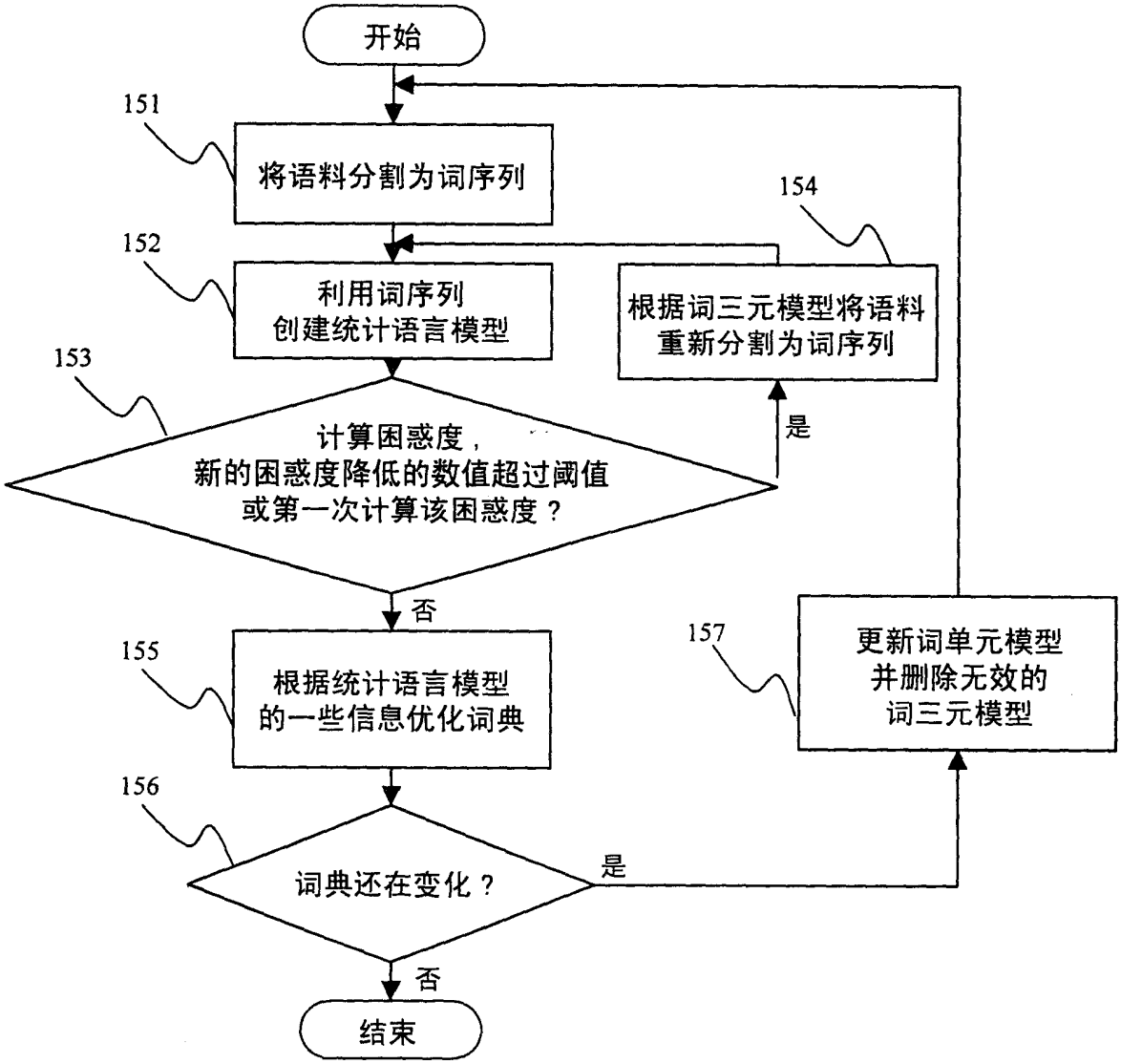

开始
151
将语料分割为词序列
152
利用词序列
创建统计语言模型
153
计算困惑度，
新的困惑度降低的数值超过阈值
或第一次计算该困惑度？
是
154
根据词三元模型将语料
重新分割为词序列
否
155
根据统计语言模型
的一些信息优化词典
156
词典还在变化？
是
157
更新词单元模型
并删除无效的
词三元模型
否
结束

图 5

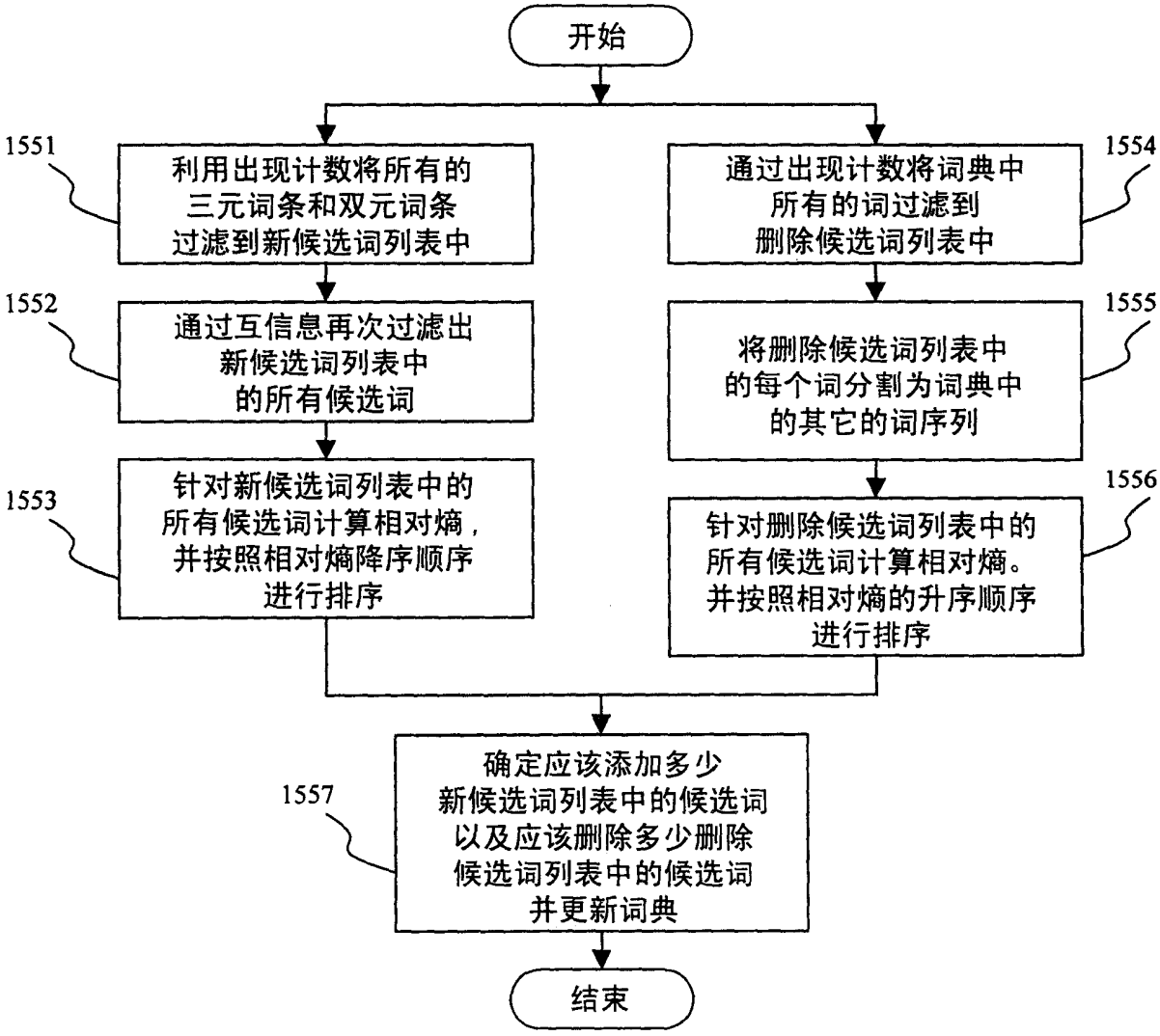

开始
1551
利用出现计数将所有的三元词条和双元词条过滤到新候选词列表中
1552
通过互信息再次过滤出新候选词列表中的所有候选词
1553
针对新候选词列表中的所有候选词计算相对熵，并按照相对熵降序顺序进行排序
1554
通过出现计数将词典中所有的词过滤到删除候选词列表中
1555
将删除候选词列表中的每个词分割为词典中的其它的词序列
1556
针对删除候选词列表中的所有候选词计算相对熵。并按照相对熵的升序顺序进行排序
1557
确定应该添加多少新候选词列表中的候选词以及应该删除多少删除候选词列表中的候选词并更新词典
结束

图 6

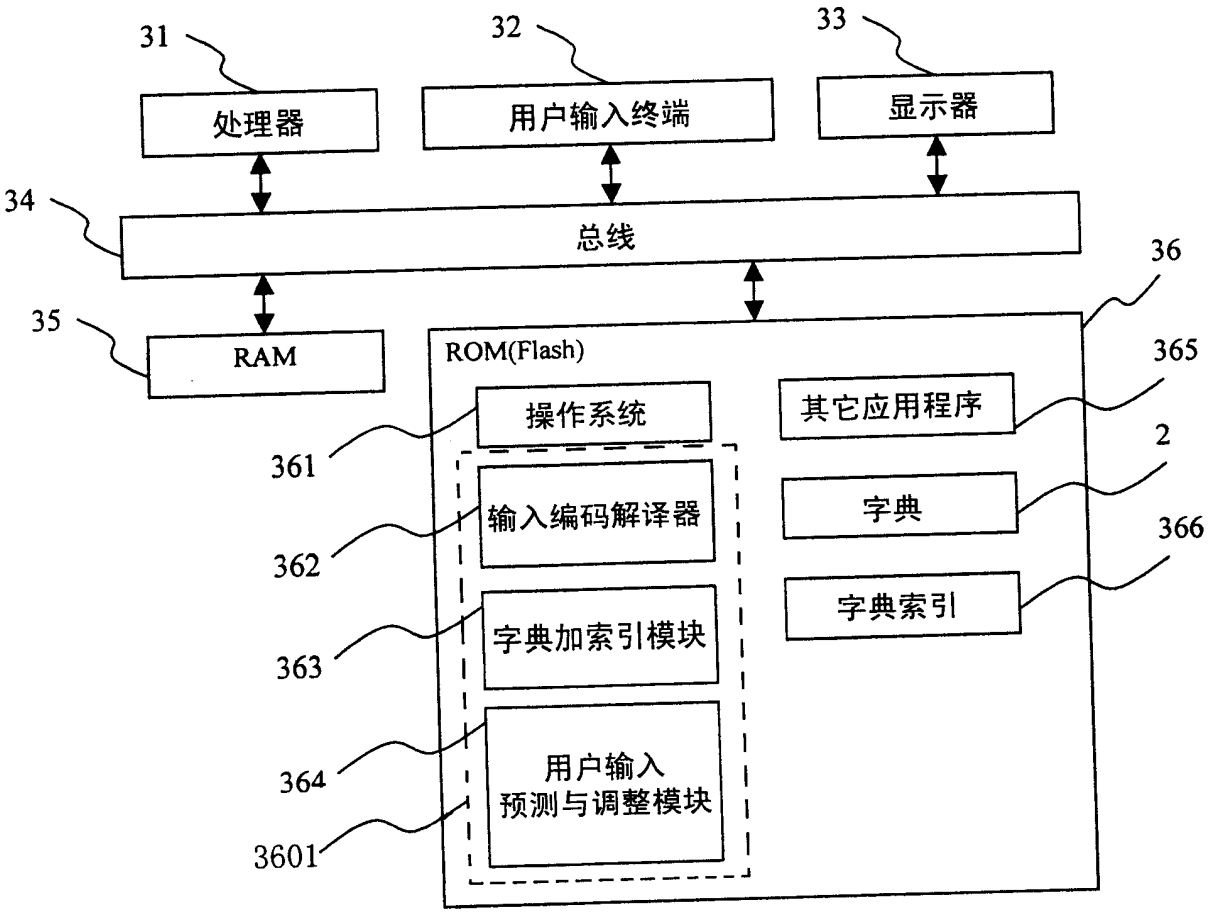

31
32
33
处理器
用户输入终端
显示器
34
总线
36
35
RAM
ROM(Flash)
365
操作系统
其它应用程序
361
2
输入编码解译器
字典
362
366
字典索引
字典加索引模块
363
364
用户输入
预测与调整模块
3601

图 7

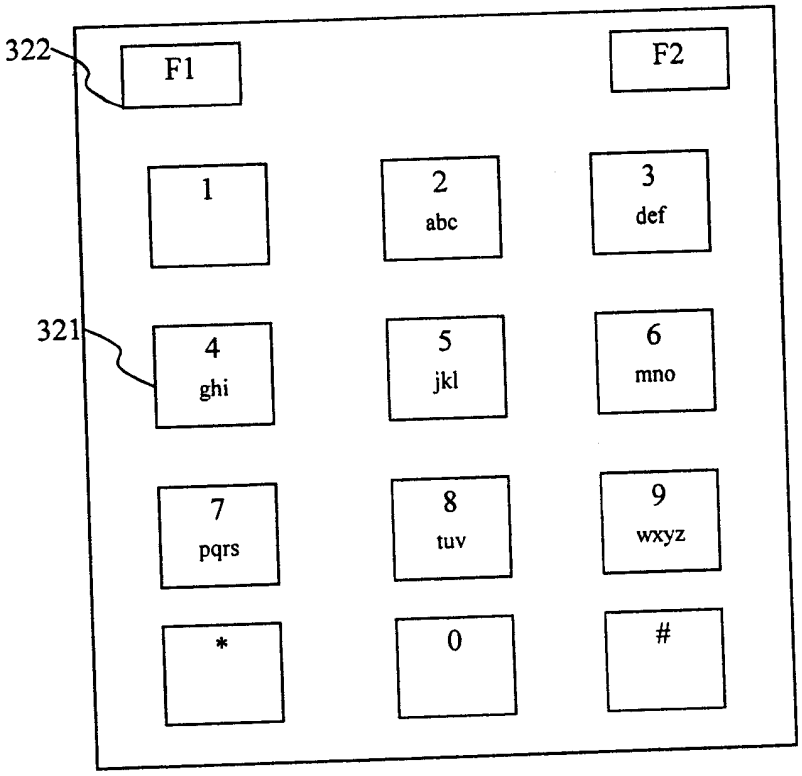

322
F1
F2
1
2
abc
3
def
321
4
ghi
5
jkl
6
mno
7
pqrs
8
tuv
9
wxyz
*
0
#

图 8A

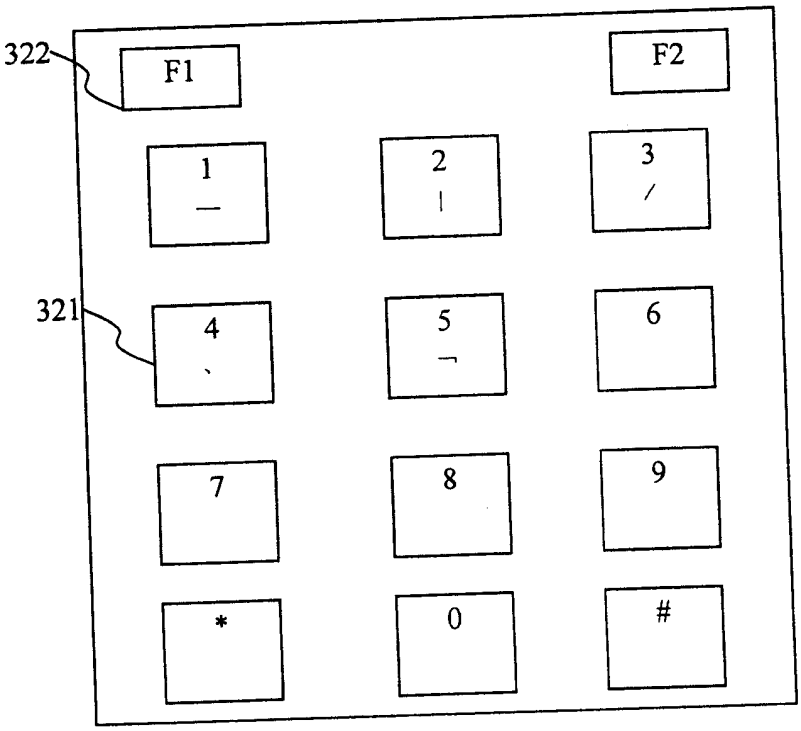

322
F1
F2
1
—
2
|
3
/
321
4
、
5
¬
6
7
8
9
*
0
#

图 8B

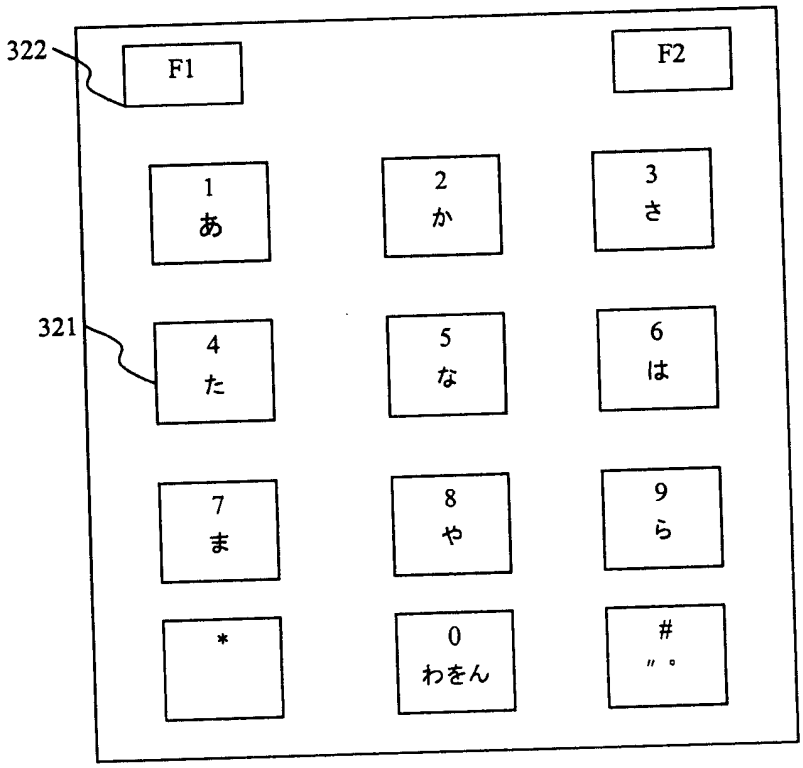


图 8C

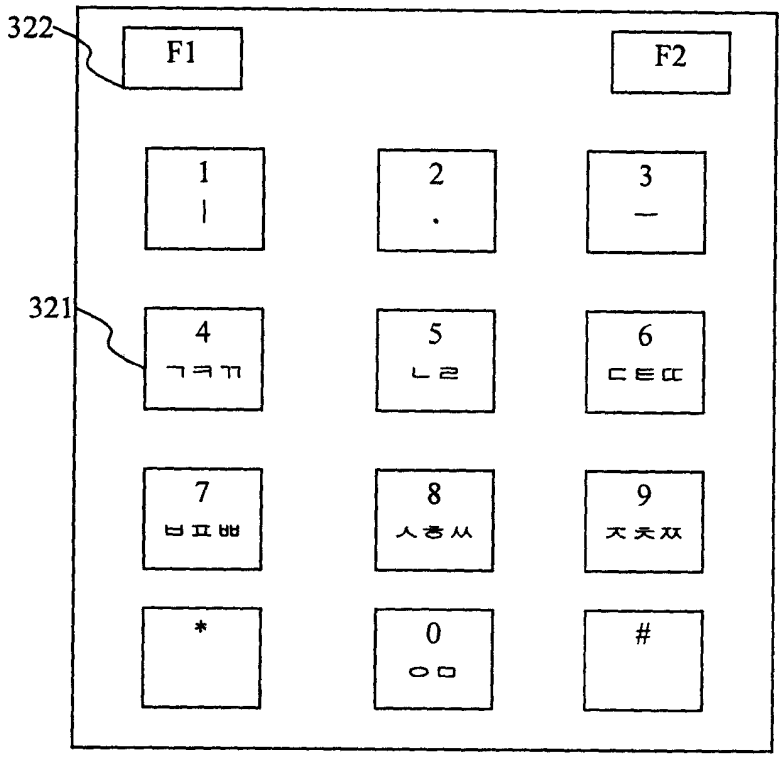


图 8D

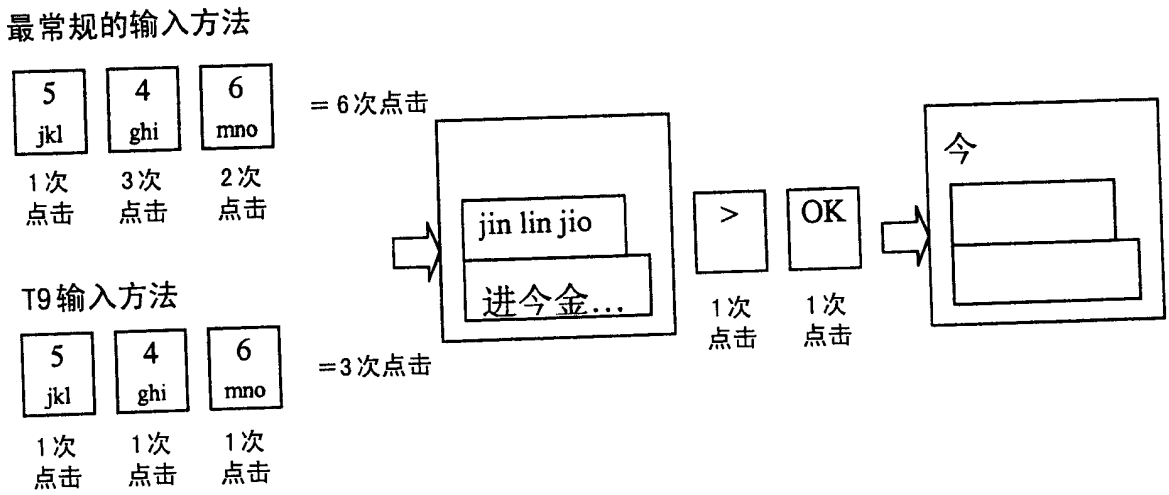

最常规的输入方法
5 jkl
4 ghi
6 mno
1次 点击
3次 点击
2次 点击
=6次点击
jin lin jio
进今金...
>
OK
1次 点击
1次 点击
今
T9输入方法
5 jkl
4 ghi
6 mno
1次 点击
1次 点击
1次 点击
=3次点击

图 9A

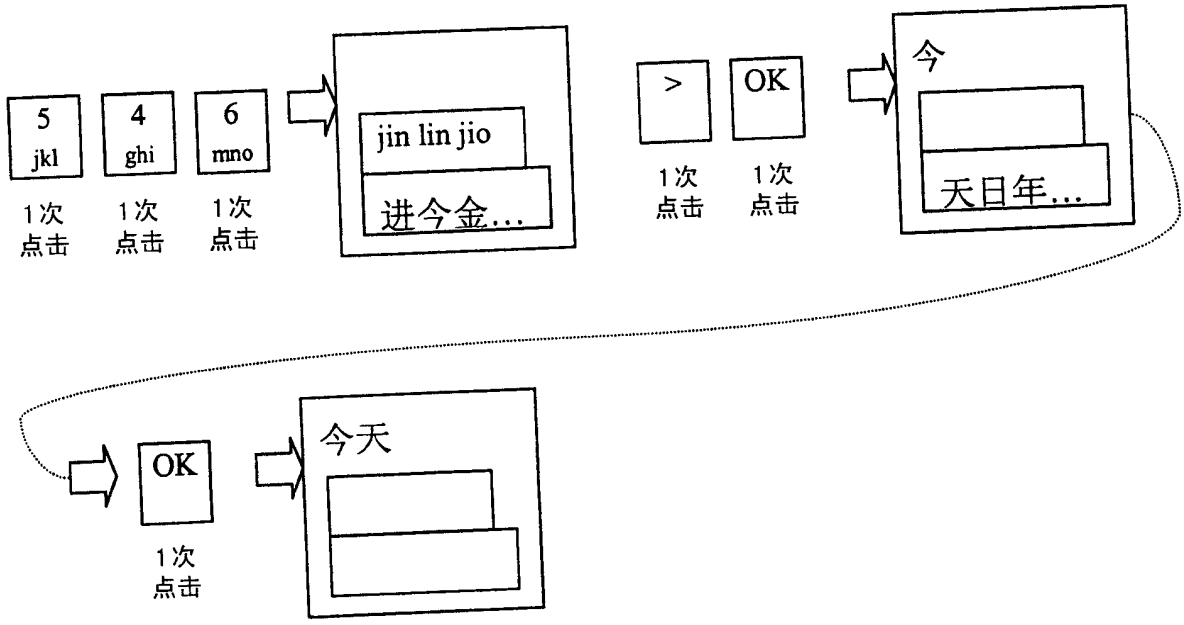

5 jkl
4 ghi
6 mno
1次 点击
1次 点击
1次 点击
jin lin jio
进今金...
>
OK
1次 点击
1次 点击
今
天日年...
OK
1次 点击
今天

图 9B

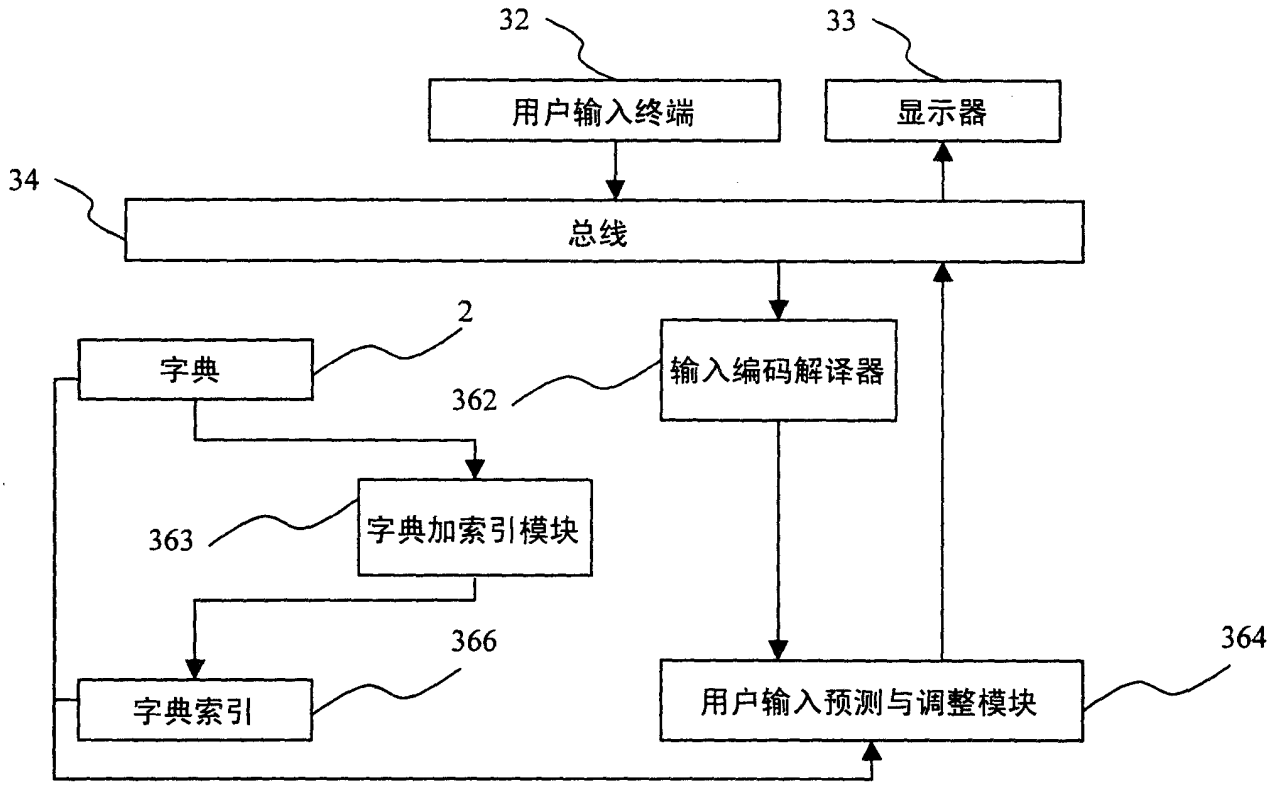

32
33
用户输入终端
显示器
34
总线
2
字典
362
输入编码解译器
363
字典加索引模块
366
字典索引
364
用户输入预测与调整模块

图 10

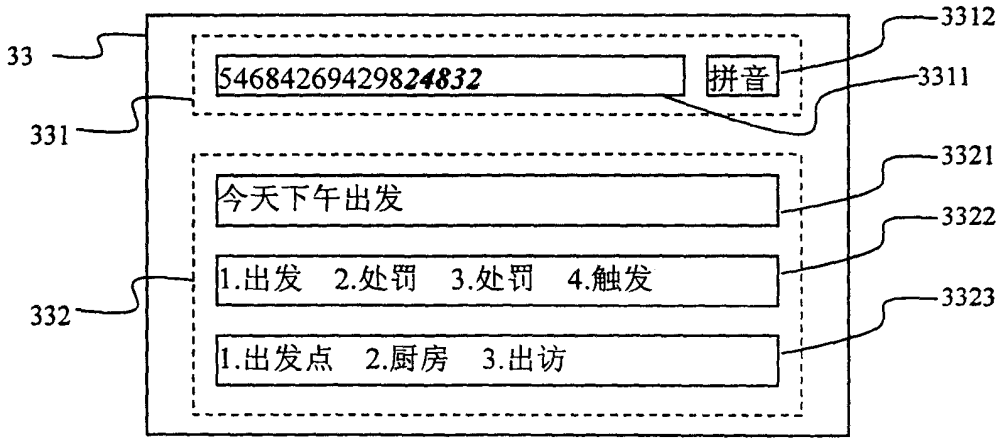

33
3312
5468426942982*4832*
拼音
3311
331
3321
今天下午出发
3322
1.出发 2.处罚 3.处罚 4.触发
332
3323
1.出发点 2.厨房 3.出访

图 11

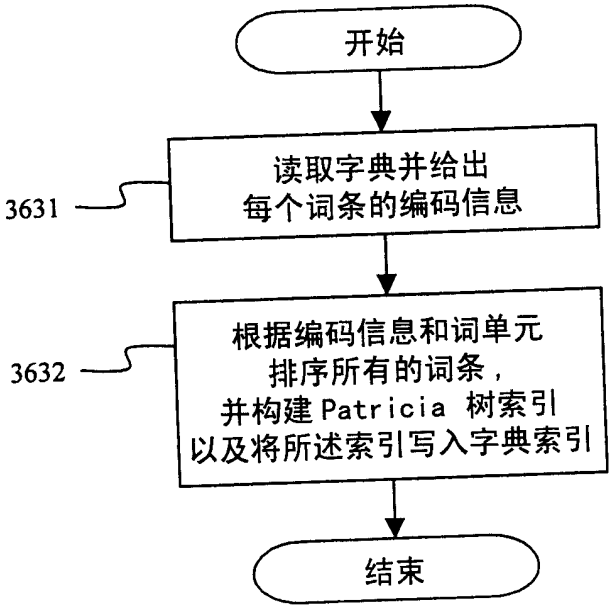

开始
读取字典并给出
每个词条的编码信息
3631
根据编码信息和词单元
排序所有的词条，
并构建 Patricia 树索引
以及将所述索引写入字典索引
3632
结束

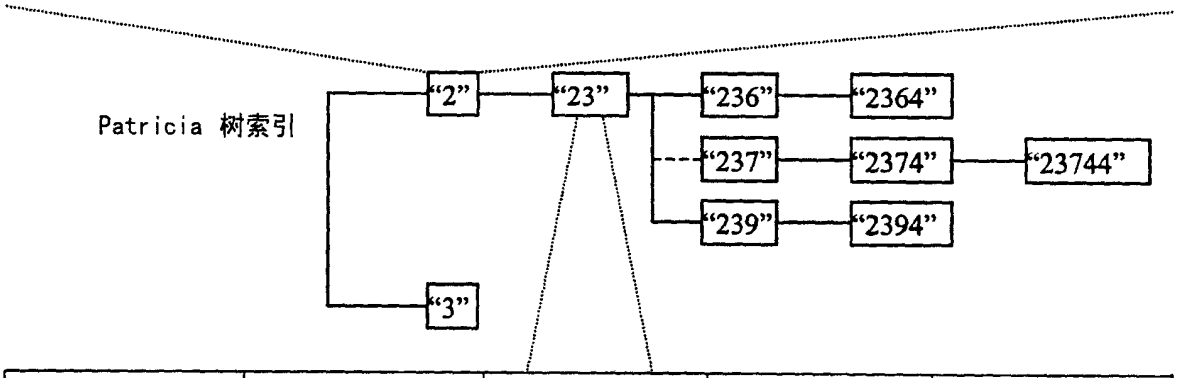
图 12

排序结果

词索引	词干	编码信息 （根据第一示例 的用户输入终端）	词单元
1	阿	“2”	4.08e-4
2	啊	“2”	1.18e-4
3	测	“23”	1.60e-4
4	笨	“236”	1.22e-5
5	层	“2364”	5.38e-4
6	测试	“23744”	2.45e-4
7	侧翼	“2394”	8.95e-6
8	恶	“3”	7.41e-5

示例节点

该节点的最后数字	起始词索引	结尾词索引	下一节点的 起始索引	下一节点的 结尾索引
“2”	1 (阿)	2 (啊)	1 (“23”)	1 (“23”)



该节点的最后数字	起始词索引	结尾词索引	下一节点的 起始索引	下一节点的 结尾索引
“3”	3 (测)	3 (测)	1 (“236”)	3 (“239”)

图 13

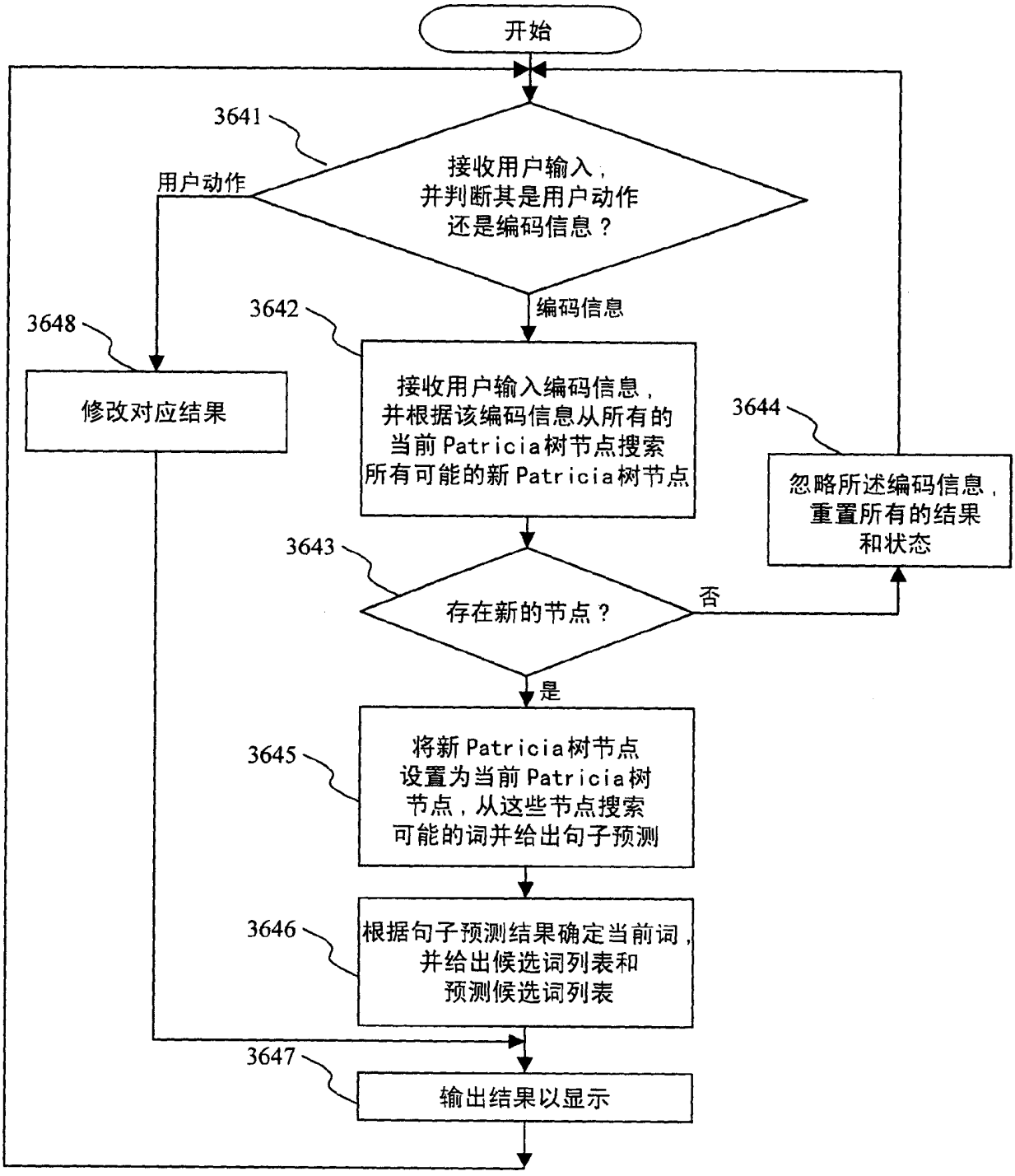

开始
3641
接收用户输入，
并判断其是用户动作
还是编码信息？
用户动作
编码信息
3642
接收用户输入编码信息，
并根据该编码信息从所有的
当前Patricia树节点搜索
所有可能的新Patricia树节点
3643
存在新的节点？
否
是
3644
忽略所述编码信息，
重置所有的结果
和状态
3648
修改对应结果
3645
将新Patricia树节点
设置为当前Patricia树
节点，从这些节点搜索
可能的词并给出句子预测
3646
根据句子预测结果确定当前词，
并给出候选词列表和
预测候选词列表
3647
输出结果以显示

图 14

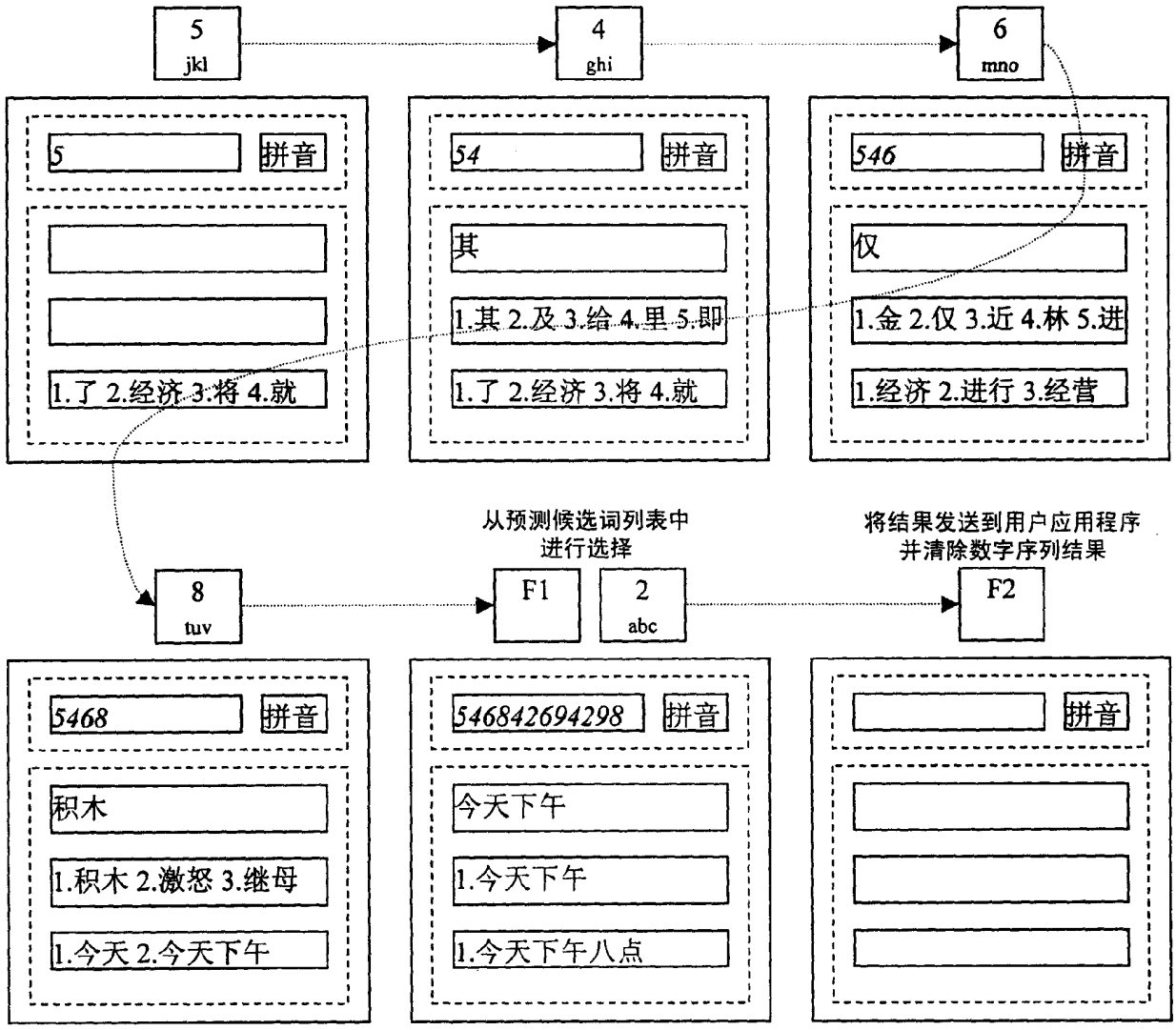

5
jkl
4
ghi
6
mno
5
拼音
54
拼音
546
拼音
其
1.其 2.及 3.给 4.里 5.即
仅
1.金 2.仅 3.近 4.林 5.进
1.了 2.经济 3.将 4.就
1.了 2.经济 3.将 4.就
1.经济 2.进行 3.经营
从预测候选词列表中
进行选择
将结果发送到用户应用程序
并清除数字序列结果
8
tuv
F1
2
abc
F2
5468
拼音
546842694298
拼音
拼音
积木
今天下午
1.积木 2.激怒 3.继母
1.今天下午
1.今天 2.今天下午
1.今天下午八点

图 15

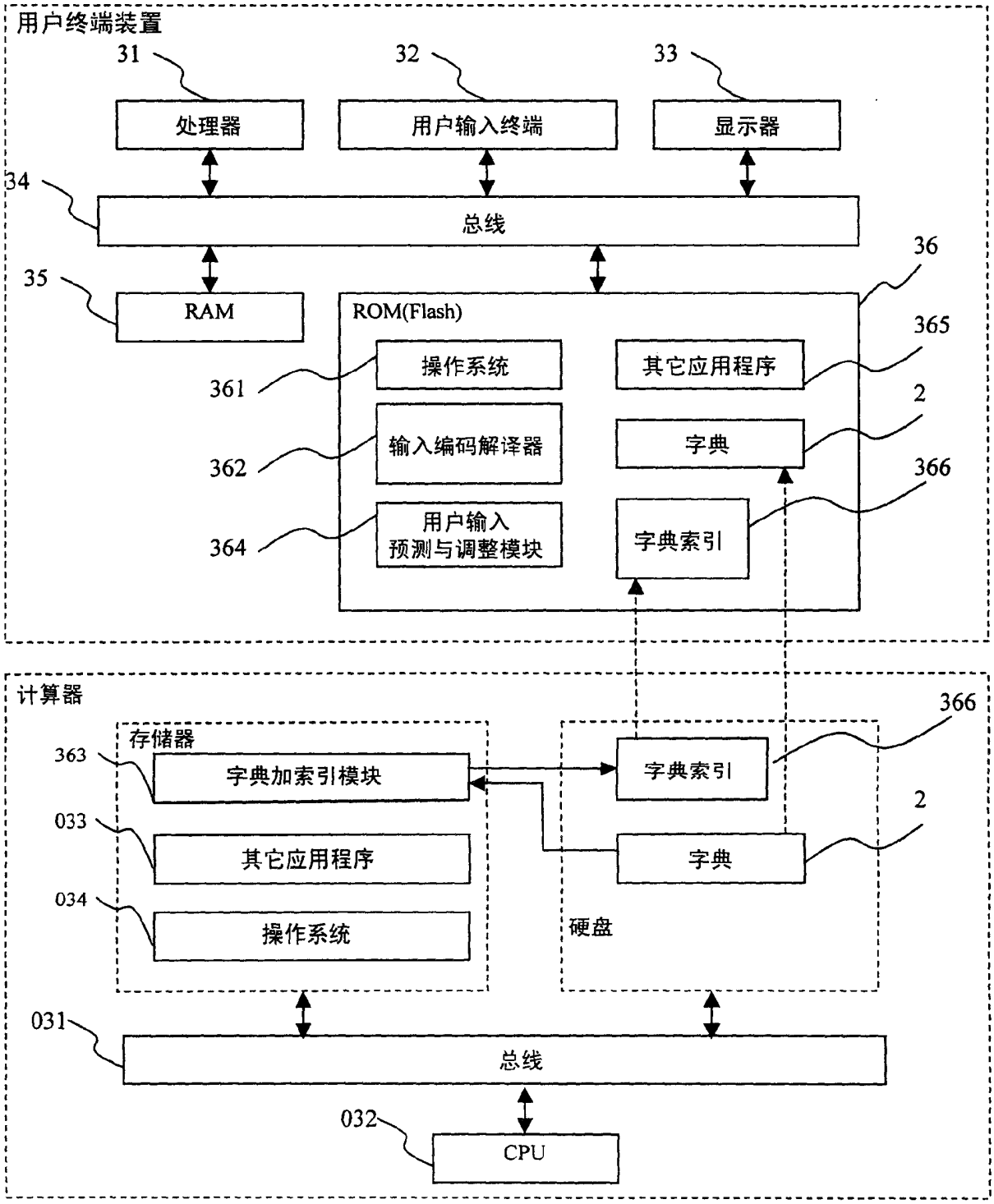

用户终端装置
31
32
33
处理器
用户输入终端
显示器
34
总线
36
35
RAM
ROM(Flash)
365
操作系统
其它应用程序
361
2
输入编码解译器
字典
362
366
用户输入
预测与调整模块
字典索引
364
计算器
366
存储器
363
字典加索引模块
字典索引
033
2
其它应用程序
字典
034
操作系统
硬盘
031
总线
032
CPU

图 16