(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106527757 A
(43)申请公布日 2017.03.22

(21)申请号 201610970625.X

(22)申请日 2016.10.28

(71)申请人 上海智臻智能网络科技股份有限公司
地址 201803 上海市嘉定区金沙江西路1555弄398号7层

(72)发明人 陈培华 朱频频 陈成才

(74)专利代理机构 工业和信息化部电子专利中心 11010
代理人 罗丹

(51)Int.Cl.
G06F 3/023(2006.01)
G06F 17/27(2006.01)

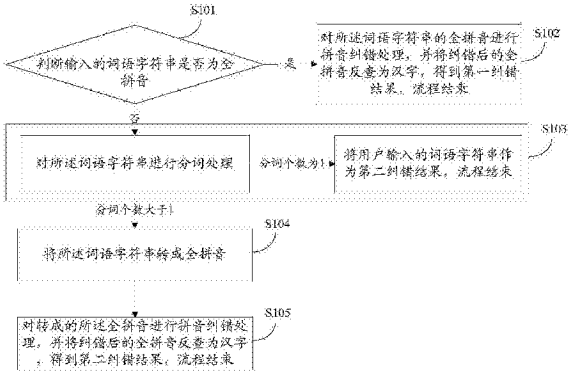
权利要求书4页 说明书16页 附图5页

(54)发明名称

一种输入纠错方法及装置

(57)摘要

本发明提出了一种输入纠错方法及装置，该方法包括：判断输入的词语字符串是否为全拼音；若是，则对所述词语字符串的全拼音进行拼音纠错处理，并将纠错后的全拼音反查为汉字，得到第一纠错结果；否则，对所述词语字符串进行分词处理，在分词处理的结果中分词个数大于1时，将所述词语字符串转成全拼音，并对转成的所述全拼音进行拼音纠错处理，并将纠错后的全拼音反查为汉字，得到第二纠错结果。本发明巧妙地将相似度计算方法应用于拼音字符的相似度计算和中文字符的相似度计算中，将本发明应用于中文搜索引擎和智能问答系统中，可以显著提高中文搜索引擎和智能问答系统中针对词语输入的信息查询和问答的准确率。



CN 106527757 A

1.一种输入纠错方法,其特征在于,包括:

判断输入的词语字符串是否为全拼音;

若是,则对所述词语字符串的全拼音进行拼音纠错处理,并将纠错后的全拼音反查为汉字,得到第一纠错结果;否则,对所述词语字符串进行分词处理,在分词处理的结果中分词个数大于1时,将所述词语字符串转成全拼音,并对转成的所述全拼音进行拼音纠错处理,并将纠错后的全拼音反查为汉字,得到第二纠错结果。

2.根据权利要求1所述的输入纠错方法,其特征在于,根据拼音反查表将纠错后的全拼音反查为汉字;所述方法还包括:预先建立拼音反查表,包括:

提供训练语料;

对训练语料进行分词以得到词语列表;

在词语列表的基础上利用拼音反查表生成工具生成拼音反查表。

3.根据权利要求1所述的输入纠错方法,其特征在于,在分词处理的结果中分词个数大于1时,所述将所述词语字符串转成全拼音,包括:

在不改变所述词语字符串中各分词出现顺序的情况下,将所述词语字符串中的汉字分词转换成拼音,再与所述词语字符串中已有的拼音一起,组成所述词语字符串对应的全拼音。

4.根据权利要求2所述的输入纠错方法,其特征在于,所述方法还包括:预先建立词频表,包括:

提供训练语料;

对训练语料进行分词以得到词语列表;

采用统计的方式根据词语列表得到词频表;

对任一词语字符串的全拼音进行拼音纠错处理,并将纠错后的全拼音反查为汉字,得到纠错结果,包括:

根据拼音反查表判断所述任一词语字符串的全拼音的拼写是否正确;

若是,则根据所述任一词语字符串的全拼音获取同音的词语列表,基于获取的同音的词语列表得到纠错结果;

若否,则根据所述任一词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度,确定出纠错结果。

5.根据权利要求4所述的输入纠错方法,其特征在于,所述基于获取的同音的词语列表得到纠错结果,包括:

判断获取的同音的词语列表是否为空,若是,则得到的纠错结果为空,否则将获取的同音的词语列表中的词语作为纠错结果;

所述根据所述词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度,确定出纠错结果,包括:

依次计算所述词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度;

对于计算出的相似度大于设定相似度阈值的情况,将拼音反查表中的拼音对应的词语列表中的词语作为纠错结果。

6.根据权利要求4所述的输入纠错方法,其特征在于,所述方法,还包括:

若分词处理的结果中分词的个数为1,则将用户输入的词语字符串作为第二纠错结果。

7.根据权利要求6所述的输入纠错方法，其特征在于，所述方法还包括：

根据第一纠错结果或者第二纠错结果进行相应的提示。

8.根据权利要求7所述的输入纠错方法，其特征在于，根据第一纠错结果进行相应的提示，包括：

判断第一纠错结果是否为空，若是，则提示用户所输入的词语字符串所对应的汉字词语数目超过一个或者用户所输入的词语字符串有误，否则将纠错结果中的词语按照在词频表中的词频从大到小的排列输出设定个数的词语以提示给用户。

9.根据权利要求7所述的输入纠错方法，其特征在于，根据第二纠错结果进行相应的提示，包括：

若第二纠错结果为空，则提示用户所输入的词语字符串所对应的汉字词语数目超过一个或者用户所输入的词语字符串中的拼音有误；

若第二纠错结果中词语的个数为1，则将第二纠错结果中的词语输出以提示给用户；

若第二纠错结果中词语的个数大于1，则根据第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度确定将第二纠错结果中的各词语向用户进行提示的方式，并进行提示。

10.根据权利要求9所述的输入纠错方法，其特征在于，所述根据第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度确定将第二纠错结果中的各词语向用户进行提示的方式，并进行提示，包括：

分别计算第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度；

若计算出的第二纠错结果中各词语对应的所述相似度数值均一致，则将第二纠错结果中的各词语按照在词频表中的词频从大到小的排列输出设定个数的词语以提示给用户，否则将第二纠错结果中的各词语按照相似度从大到小输出设定个数的词语以提示给用户。

11.根据权利要求9或10所述的输入纠错方法，其特征在于，所述输入的词语字符串中的汉字词语字符串的获取过程包括：在分词处理的结果中分词个数大于1的情况下，依次记录所述输入的词语字符串的分词中的所有汉字分词并组成汉字词语字符串。

12.一种输入纠错装置，其特征在于，包括：

判断模块，用于判断输入的词语字符串是否为全拼音；若是，则将所述词语字符串的全拼音发送给纠错模块进行处理，得到第一纠错结果；否则，将所述词语字符串发送给分词模块进行分词处理；

分词模块，用于对判断模块发来的词语字符串进行分词处理，在分词处理的结果中分词个数大于1时，将所述词语字符串发送给转换模块；

转换模块，用于将分词模块发来的词语字符串转成全拼音并将转成的全拼音发送给纠错模块进行处理，得到第二纠错结果；

纠错模块，用于对判断模块或者转换模块发来的全拼音进行拼音纠错处理，并将纠错后的全拼音反查为汉字，得到相应的纠错结果。

13.根据权利要求12所述的输入纠错装置，其特征在于，所述纠错模块，具体用于：根据拼音反查表将纠错后的全拼音反查为汉字；

所述装置还包括：

建立模块，用于：提供训练语料；对训练语料进行分词以得到词语列表；在词语列表的基础上利用拼音反查表生成工具生成拼音反查表。

14.根据权利要求12所述的输入纠错装置，其特征在于，所述转换模块，具体用于：

在不改变所述词语字符串中各分词出现顺序的情况下，将所述词语字符串中的汉字分词转换成拼音，再与所述词语字符串中已有的拼音一起，组成所述词语字符串对应的全拼音。

15.根据权利要求13所述的输入纠错装置，其特征在于，所述建立模块，还用于：采用统计的方式根据词语列表得到词频表；

所述纠错模块，包括：

拼写检查单元，用于对任一词语字符串的全拼音进行拼音纠错处理时，根据拼音反查表判断所述任一词语字符串的全拼音的拼写是否正确；若是，则调用第一处理单元，否则调用第二处理单元；

第一处理单元，用于根据所述任一词语字符串的全拼音获取同音的词语列表，基于获取的同音的词语列表得到纠错结果；

第二处理单元，用于根据所述任一词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度，确定出纠错结果。

16.根据权利要求15所述的输入纠错装置，其特征在于，所述第一处理单元，具体用于：

判断获取的同音的词语列表是否为空，若是，则得到的纠错结果为空，否则将获取的同音的词语列表中的词语作为纠错结果；

所述第二处理单元，具体用于：

依次计算所述词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度；对于计算出的相似度大于设定相似度阈值的情况，将拼音反查表中的拼音对应的词语列表中的词语作为纠错结果。

17.根据权利要求15所述的输入纠错装置，其特征在于，所述分词模块，还用于：若分词处理的结果中分词的个数为1，则将用户输入的词语字符串作为第二纠错结果。

18.根据权利要求17所述的输入纠错装置，其特征在于，所述装置还包括：

提示模块，用于根据第一纠错结果或者第二纠错结果进行相应的提示。

19.根据权利要求18所述的输入纠错装置，其特征在于，所述提示模块，具体用于：

判断第一纠错结果是否为空，若是，则提示用户所输入的词语字符串所对应的汉字词语数目超过一个或者用户所输入的词语字符串有误，否则将纠错结果中的词语按照在词频表中的词频从大到小的排列输出设定个数的词语以提示给用户。

20.根据权利要求18所述的输入纠错装置，其特征在于，所述提示模块，具体用于：

若第二纠错结果为空，则提示用户所输入的词语字符串所对应的汉字词语数目超过一个或者用户所输入的词语字符串中的拼音有误；

若第二纠错结果中词语的个数为1，则将第二纠错结果中的词语输出以提示给用户；

若第二纠错结果中词语的个数大于1，则根据第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度确定将第二纠错结果中的各词语向用户进行提示的方式，并进行提示。

21.根据权利要求20所述的输入纠错装置，其特征在于，所述提示模块，在根据第二纠

错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度确定将第二纠错结果中的各词语向用户进行提示的方式时，具体用于：

分别计算第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度；

若计算出的第二纠错结果中各词语对应的所述相似度数值均一致，则将第二纠错结果中的各词语按照在词频表中的词频从大到小的排列输出设定个数的词语以提示给用户，否则将第二纠错结果中的各词语按照相似度从大到小输出设定个数的词语以提示给用户。

22.根据权利要求20或21所述的输入纠错装置，其特征在于，所述分词模块，还用于：在分词处理的结果中分词个数大于1的情况下，依次记录所述输入的词语字符串的分词中的所有汉字分词并组成汉字词语字符串，发送给所述提示模块。

一种输入纠错方法及装置

技术领域

[0001] 本发明涉及自然语音处理和机器学习技术领域，尤其涉及一种输入纠错方法及装置。

背景技术

[0002] 目前用户常常会通过中文搜索引擎或智能问答系统进行信息查询，其中很大一部分查询是以词语的形式输入的。以百度为代表的中文搜索引擎和以小i机器人为代表的智能问答系统均能对用户输入的中文词语进行相应的响应和反馈。但是，当用户输入错误的词语时，主要有：同音别字、近音别字、形近别字、拼音、多字漏字等情况，以上搜索引擎或智能问答系统就可能会无法正确或有效处理此类词语，致使用户无法获取需要的信息。例如，原词为：火中取栗，对于存在同音别字、近音别字、拼音等错误的“火宗去li”或存在同音别字、多字漏字等错误的“火中去”，以上的搜索引擎或智能问答系统均无法正确处理。

发明内容

[0003] 本发明要解决的技术问题是，提供一种输入纠错方法及装置，对输入的同音别字、近音别字、拼音、形近别字、多字漏字等情况进行有效的纠错处理。

[0004] 本发明采用的技术方案是，所述输入纠错方法，包括：

[0005] 判断输入的词语字符串是否为全拼音；

[0006] 若是，则对所述词语字符串的全拼音进行拼音纠错处理，并将纠错后的全拼音反查为汉字，得到第一纠错结果；否则，对所述词语字符串进行分词处理，在分词处理的结果中分词个数大于1时，将所述词语字符串转成全拼音，并对转成的所述全拼音进行拼音纠错处理，并将纠错后的全拼音反查为汉字，得到第二纠错结果。

[0007] 进一步的，根据拼音反查表将纠错后的全拼音反查为汉字；

[0008] 所述方法还包括：预先建立拼音反查表，包括：

[0009] 提供训练语料；

[0010] 对训练语料进行分词以得到词语列表；

[0011] 在词语列表的基础上利用拼音反查表生成工具生成拼音反查表。

[0012] 进一步的，在分词处理的结果中分词个数大于1时，所述将所述词语字符串转成全拼音，包括：

[0013] 在不改变所述词语字符串中各分词出现顺序的情况下，将所述词语字符串中的汉字分词转换成拼音，再与所述词语字符串中已有的拼音一起，组成所述词语字符串对应的全拼音。

[0014] 进一步的，所述方法还包括：预先建立词频表，包括：

[0015] 提供训练语料；

[0016] 对训练语料进行分词以得到词语列表；

[0017] 采用统计的方式根据词语列表得到词频表；

[0018] 对任一词语字符串的全拼音进行拼音纠错处理，并将纠错后的全拼音反查为汉字，得到纠错结果，包括：

[0019] 根据拼音反查表判断所述任一词语字符串的全拼音的拼写是否正确；

[0020] 若是，则根据所述任一词语字符串的全拼音获取同音的词语列表，基于获取的同音的词语列表得到纠错结果；

[0021] 若否，则根据所述任一词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度，确定出纠错结果。

[0022] 进一步的，所述基于获取的同音的词语列表得到纠错结果，包括：

[0023] 判断获取的同音的词语列表是否为空，若是，则得到的纠错结果为空，否则将获取的同音的词语列表中的词语作为纠错结果；

[0024] 所述根据所述词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度，确定出纠错结果，包括：

[0025] 依次计算所述词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度；

[0026] 对于计算出的相似度大于设定相似度阈值的情况，将拼音反查表中的拼音对应的词语列表中的词语作为纠错结果。

[0027] 进一步的，所述方法，还包括：

[0028] 若分词处理的结果中分词的个数为1，则将用户输入的词语字符串作为第二纠错结果。

[0029] 进一步的，所述方法还包括：

[0030] 根据第一纠错结果或者第二纠错结果进行相应的提示。

[0031] 进一步的，根据第一纠错结果进行相应的提示，包括：

[0032] 判断第一纠错结果是否为空，若是，则提示用户所输入的词语字符串所对应的汉字词语数目超过一个或者用户所输入的词语字符串有误，否则将纠错结果中的词语按照在词频表中的词频从大到小的排列输出设定个数的词语以提示给用户。

[0033] 进一步的，根据第二纠错结果进行相应的提示，包括：

[0034] 若第二纠错结果为空，则提示用户所输入的词语字符串所对应的汉字词语数目超过一个或者用户所输入的词语字符串中的拼音有误；

[0035] 若第二纠错结果中词语的个数为1，则将第二纠错结果中的词语输出以提示给用户；

[0036] 若第二纠错结果中词语的个数大于1，则根据第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度确定将第二纠错结果中的各词语向用户进行提示的方式，并进行提示。

[0037] 进一步的，所述根据第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度确定将第二纠错结果中的各词语向用户进行提示的方式，并进行提示，包括：

[0038] 分别计算第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度；

[0039] 若计算出的第二纠错结果中各词语对应的所述相似度数值均一致，则将第二纠错结果中的各词语按照在词频表中的词频从大到小的排列输出设定个数的词语以提示给用

户，否则将第二纠错结果中的各词语按照相似度从大到小输出设定个数的词语以提示给用户。

[0040] 进一步的，所述输入的词语字符串中的汉字词语字符串的获取过程包括：在分词处理的结果中分词个数大于1的情况下，依次记录所述输入的词语字符串的分词中的所有汉字分词并组成汉字词语字符串。

[0041] 本发明还提供一种输入纠错装置，包括：

[0042] 判断模块，用于判断输入的词语字符串是否为全拼音；若是，则将所述词语字符串的全拼音发送给纠错模块进行处理，得到第一纠错结果；否则，将所述词语字符串发送给分词模块进行分词处理；

[0043] 分词模块，用于对判断模块发来的词语字符串进行分词处理，在分词处理的结果中分词个数大于1时，将所述词语字符串发送给转换模块；

[0044] 转换模块，用于将分词模块发来的词语字符串转成全拼音并将转成的全拼音发送给纠错模块进行处理，得到第二纠错结果；

[0045] 纠错模块，用于对判断模块或者转换模块发来的全拼音进行拼音纠错处理，并将纠错后的全拼音反查为汉字，得到相应的纠错结果。

[0046] 进一步的，所述纠错模块，具体用于：根据拼音反查表将纠错后的全拼音反查为汉字；

[0047] 所述装置还包括：

[0048] 建立模块，用于：提供训练语料；对训练语料进行分词以得到词语列表；在词语列表的基础上利用拼音反查表生成工具生成拼音反查表。

[0049] 进一步的，所述转换模块，具体用于：

[0050] 在不改变所述词语字符串中各分词出现顺序的情况下，将所述词语字符串中的汉字分词转换成拼音，再与所述词语字符串中已有的拼音一起，组成所述词语字符串对应的全拼音。

[0051] 进一步的，所述建立模块，还用于：采用统计的方式根据词语列表得到词频表；

[0052] 所述纠错模块，包括：

[0053] 拼写检查单元，用于对任一词语字符串的全拼音进行拼音纠错处理时，根据拼音反查表判断所述任一词语字符串的全拼音的拼写是否正确；若是，则调用第一处理单元，否则调用第二处理单元；

[0054] 第一处理单元，用于根据所述任一词语字符串的全拼音获取同音的词语列表，基于获取的同音的词语列表得到纠错结果；

[0055] 第二处理单元，用于根据所述任一词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度，确定出纠错结果。

[0056] 进一步的，所述第一处理单元，具体用于：

[0057] 判断获取的同音的词语列表是否为空，若是，则得到的纠错结果为空，否则将获取的同音的词语列表中的词语作为纠错结果；

[0058] 所述第二处理单元，具体用于：

[0059] 依次计算所述词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度；对于计算出的相似度大于设定相似度阈值的情况，将拼音反查表中的拼音对应的词语列表中

的词语作为纠错结果。

[0060] 进一步的，所述分词模块，还用于：若分词处理的结果中分词的个数为1，则将用户输入的词语字符串作为第二纠错结果。

[0061] 进一步的，所述装置还包括：

[0062] 提示模块，用于根据第一纠错结果或者第二纠错结果进行相应的提示。

[0063] 进一步的，所述提示模块，具体用于：

[0064] 判断第一纠错结果是否为空，若是，则提示用户所输入的词语字符串所对应的汉字词语数目超过一个或者用户所输入的词语字符串有误，否则将纠错结果中的词语按照在词频表中的词频从大到小的排列输出设定个数的词语以提示给用户。

[0065] 进一步的，所述提示模块，具体用于：

[0066] 若第二纠错结果为空，则提示用户所输入的词语字符串所对应的汉字词语数目超过一个或者用户所输入的词语字符串中的拼音有误；

[0067] 若第二纠错结果中词语的个数为1，则将第二纠错结果中的词语输出以提示给用户；

[0068] 若第二纠错结果中词语的个数大于1，则根据第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度确定将第二纠错结果中的各词语向用户进行提示的方式，并进行提示。

[0069] 进一步的，所述提示模块，在根据第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度确定将第二纠错结果中的各词语向用户进行提示的方式时，具体用于：

[0070] 分别计算第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度；

[0071] 若计算出的第二纠错结果中各词语对应的所述相似度数值均一致，则将第二纠错结果中的各词语按照在词频表中的词频从大到小的排列输出设定个数的词语以提示给用户，否则将第二纠错结果中的各词语按照相似度从大到小输出设定个数的词语以提示给用户。

[0072] 进一步的，所述分词模块，还用于：在分词处理的结果中分词个数大于1的情况下，依次记录所述输入的词语字符串的分词中的所有汉字分词并组成汉字词语字符串，发送给所述提示模块。

[0073] 采用上述技术方案，本发明所述输入纠错方法及装置至少具有下列优点：

[0074] 1、本发明提供的输入纠错方法中，拼音纠错处理可有效处理同音别字、近音别字、拼音、形近别字、多字漏字等词语输入错误问题，中文字符相似度计算和拼音纠错处理相结合可以进一步提高纠错的准确性。

[0075] 2、本发明实施例中在建立词语列表、拼音反查表和词频表的过程中，充分有效的利用训练语料所提供的词语信息，能够快速适用于不同领域的自定义词语纠错。

[0076] 3、本发明巧妙地将相似度计算方法应用于拼音字符的相似度计算和中文字符的相似度计算中，并采用了不同的评价指标，分别是在进行拼音字符的相似度计算时使用的相似度阈值、以及在向用户提示纠错结果时用到相似度从大到小排列后按照设定个数进行输出，以得到比较准确的计算结果和输出结果。

[0077] 4、本发明所提供的所述输入纠错方法及装置，应用于中文搜索引擎和智能问答系统中，可以显著提高中文搜索引擎和智能问答系统中针对词语输入的信息查询和问答的准确率。

附图说明

[0078] 图1为本发明第一实施例的输入纠错方法流程图；

[0079] 图2为本发明第二实施例的输入纠错方法流程图；

[0080] 图3为本发明第三实施例的输入纠错方法流程图；

[0081] 图4为本发明第四实施例的输入纠错装置组成结构示意图；

[0082] 图5为本发明第五实施例的输入纠错装置组成结构示意图；

[0083] 图6为本发明第六实施例的输入纠错装置组成结构示意图；

[0084] 图7为本发明第七实施例的基于分词和相似度计算的输入纠错方法流程图；

[0085] 图8为本发明第七实施例的基于分词和相似度计算的输入纠错系统示意图。

具体实施方式

[0086] 为更进一步阐述本发明为达成预定目的所采取的技术手段及功效，以下结合附图及较佳实施例，对本发明进行详细说明如后。

[0087] 本发明第一实施例，一种输入纠错方法，如图1所示，包括以下具体步骤：

[0088] 步骤S101，判断输入的词语字符串是否为全拼音；若是，则执行步骤S102，否则，执行步骤S103。

[0089] 步骤S102，对所述词语字符串的全拼音进行拼音纠错处理，并将纠错后的全拼音反查为汉字，得到第一纠错结果，流程结束。

[0090] 步骤S103，对所述词语字符串进行分词处理，在分词处理的结果中分词个数大于1时，执行步骤S104；若分词处理的结果中分词的个数为1，则将用户输入的词语字符串作为第二纠错结果，流程结束。

[0091] 步骤S104，将所述词语字符串转成全拼音。

[0092] 具体的，步骤S104包括：

[0093] 在不改变所述词语字符串中各分词出现顺序的情况下，将所述词语字符串中的汉字分词转换成拼音，再与所述词语字符串中已有的拼音一起，组成所述词语字符串对应的全拼音。

[0094] 步骤S105，对转成的所述全拼音进行拼音纠错处理，并将纠错后的全拼音反查为汉字，得到第二纠错结果，流程结束。

[0095] 本发明实施例的所述输入纠错方法中，通过拼音纠错处理，可有效处理出现同音别字、近音别字、拼音、形近别字、多字漏字等词语输入错误问题，将本发明实施例的所述方法应用于中文搜索引擎和智能问答系统中，可以显著提高中文搜索引擎和智能问答系统中针对词语输入的信息查询和问答的准确率。

[0096] 本发明第二实施例，一种输入纠错方法，如图2所示，包括以下具体步骤：

[0097] 步骤S201，预先建立词语列表、拼音反查表和词频表。

[0098] 具体的，步骤S201包括：

[0099] 提供训练语料；

[0100] 对训练语料进行分词以得到词语列表；

[0101] 在词语列表的基础上利用拼音反查表生成工具生成拼音反查表，采用统计的方式根据词语列表得到词频表。

[0102] 本发明实施例中在建立词语列表、拼音反查表和词频表的过程中，充分有效的利用训练语料所提供的词语信息，能够快速适用于不同领域的自定义词语纠错。

[0103] 步骤S202，判断输入的词语字符串是否为全拼音；若是，则执行步骤S203，否则，执行步骤S204。

[0104] 步骤S203，根据词语列表、拼音反查表和词频表对所述词语字符串的全拼音进行拼音纠错处理，根据拼音反查表将纠错后的全拼音反查为汉字，得到第一纠错结果，流程结束。

[0105] 步骤S204，对所述词语字符串进行分词处理，在分词处理的结果中分词个数大于1时，执行步骤S205；若分词处理的结果中分词的个数为1，则将用户输入的词语字符串作为第二纠错结果，流程结束。

[0106] 步骤S205，将所述词语字符串转成全拼音。

[0107] 具体的，步骤S205包括：

[0108] 在不改变所述词语字符串中各分词出现顺序的情况下，将所述词语字符串中的汉字分词转换成拼音，再与所述词语字符串中已有的拼音一起，组成所述词语字符串对应的全拼音。

[0109] 步骤S206，根据词语列表、拼音反查表和词频表对转成的所述全拼音进行拼音纠错处理，根据拼音反查表将纠错后的全拼音反查为汉字，得到第二纠错结果，流程结束。

[0110] 具体的，在步骤S203和步骤S206中，进行拼音纠错处理、反查汉字以得到纠错结果的方式均相同，这里统一进行详细描述如下：

[0111] 对任一词语字符串的全拼音进行拼音纠错处理，并将纠错后的全拼音反查为汉字，得到纠错结果，包括：

[0112] 根据拼音反查表判断所述任一词语字符串的全拼音的拼写是否正确；具体是用所述任一词语字符串的全拼音在拼音反查表中进行比对查找，若有一致的，则表明拼写正确，否则拼写错误。

[0113] 若是，则根据所述任一词语字符串的全拼音获取同音的词语列表，基于获取的同音的词语列表得到纠错结果；

[0114] 若否，则根据所述任一词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度，确定出纠错结果。

[0115] 进一步的，所述基于获取的同音的词语列表得到纠错结果，包括：

[0116] 判断获取的同音的词语列表是否为空，若是，则得到的纠错结果为空，否则将获取的同音的词语列表中的词语作为纠错结果；

[0117] 所述根据所述词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度，确定出纠错结果，包括：

[0118] 依次计算所述词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度；对于计算出的相似度大于设定相似度阈值的情况，将拼音反查表中的拼音对应的词语列表中

的词语作为纠错结果。

[0119] 下面以正确的原词为“火中取栗”为例，介绍一下本发明实施例的所述方法的应用过程：若用户输入存在近音别字且混有拼音的词语字符串“火宗去li”，因为该词语字符串中同时包含汉字和拼音，所以判定输入的词语字符串不是全拼音，然后对“火宗去li”进行分词处理，得到的分词为“火”“宗”“去”“li”，由于分词的个数大于1，将“火宗去li”转成全拼音为“huozongquli”，再对“huozongquli”进行拼音纠错处理，具体是首选判断该全拼音拼写是否正确，由于在拼音反查表中包含的都是正确的拼音，在拼音反查表中没有找到“huozongquli”，故该全拼音拼写错误，则根据“huozongquli”与拼音反查表中的每个拼音之间的相似度中最高者“huozhongquli”，得到与“huozhongquli”对应的词语“火中取栗”，以询问用户。

[0120] 若用户输入全拼音的词语字符串“huozhongquli”，因为该词语字符串是全拼音，所以直接对该全拼音进行拼音纠错处理，具体是在确定“huozhongquli”的拼音拼写正确的情况下，根据该全拼音获取同音的词语列表，基于获取的同音的词语列表可以得到至少包含“火中取栗”的提示内容，以询问用户。

[0121] 若用户输入存在多字和近音别字的词语字符串“火中去栗了”，由于该词语字符串不是全拼音，对“火中取栗了”进行分词处理，得到的分词为“火”“中”“去”“栗”“了”，由于分词的个数大于1，将“火中去栗了”转成全拼音为“huozhongqulile”，再对“huozhongqulile”进行拼音纠错处理，具体是首选判断该全拼音拼写是否正确，由于在拼音反查表中包含的都是正确的拼音，在拼音反查表中没有找到“huozhongqulile”，故该全拼音拼写错误，则根据“huozhongqulile”与拼音反查表中的每个拼音之间的相似度中最高者“huozhongquli”，得到与“huozhongquli”对应的词语“火中取栗”，以询问用户。

[0122] 若用户输入存在漏字和近音别字的词语字符串“火中去”，由于该词语字符串不是全拼音，对“火中去”进行分词处理，得到的分词为“火”“中”“去”，由于分词的个数大于1，将“火中去”转成全拼音为“huozhongqu”，再对“huozhongqu”进行拼音纠错处理，具体是首选判断该全拼音拼写是否正确，由于在拼音反查表中包含的都是正确的拼音，在拼音反查表中没有找到“huozhongqu”，故该全拼音拼写错误，则根据“huozhongqu”与拼音反查表中的每个拼音之间的相似度中最高者“huozhongquli”，得到与“huozhongquli”对应的词语“火中取栗”，以询问用户。

[0123] 下面以正确的原词为“十二生肖”为例，介绍一下本发明实施例的所述方法的应用过程：若用户输入存在同音别字的词语字符串“使二生效”，因为该词语字符串不是全拼音，对“使二生效”进行分词处理，得到的分词为“使”“二”“生效”，由于分词的个数大于1，将“使二生效”转成全拼音为“shiershengxiao”，再对“shiershengxiao”进行拼音纠错处理，具体是确定拼音正确时，找同音词语列表，其中必然包含“十二生肖”，从而得到至少包含“十二生肖”的提示内容，以询问用户。

[0124] 若用户输入存在近音别字的词语字符串“十而僧小”，因为该词语字符串不是全拼音，对“十而僧小”进行分词处理，得到的分词为“十”“而”“僧”“小”，由于分词的个数大于1，将“十而僧小”转成全拼音为“shiersengxiao”，再对“shiersengxiao”进行拼音纠错处理，具体是首选判断该全拼音拼写是否正确，由于在拼音反查表中包含的都是正确的拼音，在拼音反查表中没有找到“shiersengxiao”，故该全拼音拼写错误，则根据“shiersengxiao”

与拼音反查表中的每个拼音之间的相似度中最高者“shiershengxiao”，得到与“shiershengxiao”对应的词语“十二生肖”，以询问用户。

[0125] 本发明实施例的所述输入纠错方法中，通过拼音纠错处理，可有效处理出现同音别字、近音别字、拼音、形近别字、多字漏字等词语输入错误问题，将本发明实施例的所述方法应用于中文搜索引擎和智能问答系统中，可以显著提高中文搜索引擎和智能问答系统中针对词语输入的信息查询和问答的准确率。

[0126] 本发明第三实施例，一种输入纠错方法，如图3所示，包括以下具体步骤：

[0127] 步骤S201，预先建立词语列表、拼音反查表和词频表。

[0128] 具体的，步骤S201包括：

[0129] 提供训练语料；

[0130] 对训练语料进行分词以得到词语列表；

[0131] 在词语列表的基础上利用拼音反查表生成工具生成拼音反查表，采用统计的方式根据词语列表得到词频表。

[0132] 步骤S202，判断输入的词语字符串是否为全拼音；若是，则执行步骤S203，否则，执行步骤S204。

[0133] 步骤S203，根据词语列表、拼音反查表和词频表对所述词语字符串的全拼音进行拼音纠错处理，根据拼音反查表将纠错后的全拼音反查为汉字，得到第一纠错结果，执行步骤S207。

[0134] 步骤S204，对所述词语字符串进行分词处理，在分词处理的结果中分词个数大于1时，执行步骤S205；若分词处理的结果中分词的个数为1，则将用户输入的词语字符串作为第二纠错结果，执行步骤S207。

[0135] 步骤S205，将所述词语字符串转成全拼音。

[0136] 具体的，步骤S205包括：

[0137] 在不改变所述词语字符串中各分词出现顺序的情况下，将所述词语字符串中的汉字分词转换成拼音，再与所述词语字符串中已有的拼音一起，组成所述词语字符串对应的全拼音。

[0138] 步骤S206，根据词语列表、拼音反查表和词频表对转成的所述全拼音进行拼音纠错处理，根据拼音反查表将纠错后的全拼音反查为汉字，得到第二纠错结果，执行步骤S207。

[0139] 具体的，在步骤S203和步骤S206中，进行拼音纠错处理、反查汉字以得到纠错结果的方式均相同，这里统一进行详细描述如下：

[0140] 对任一词语字符串的全拼音进行拼音纠错处理，并将纠错后的全拼音反查为汉字，得到纠错结果，包括：

[0141] 根据拼音反查表判断所述任一词语字符串的全拼音的拼写是否正确；实际上就是，用所述任一词语字符串的全拼音在拼音反查表中进行比对查找，若有一致的，则表明拼写正确，否则拼写错误。

[0142] 若是，则根据所述任一词语字符串的全拼音获取同音的词语列表，基于获取的同音的词语列表得到纠错结果；

[0143] 若否，则根据所述任一词语字符串的全拼音与拼音反查表中的每个拼音之间的相

似度，确定出纠错结果。

[0144] 进一步的，所述基于获取的同音的词语列表得到纠错结果，包括：

[0145] 判断获取的同音的词语列表是否为空，若是，则得到的纠错结果为空，否则将获取的同音的词语列表中的词语作为纠错结果；

[0146] 所述根据所述词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度，确定出纠错结果，包括：

[0147] 依次计算所述词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度；对于计算出的相似度大于设定相似度阈值的情况，将拼音反查表中的拼音对应的词语列表中的词语作为纠错结果。

[0148] 步骤S207，根据第一纠错结果或者第二纠错结果进行相应的提示。

[0149] 具体的，在步骤S207中，根据第一纠错结果进行相应的提示，包括：

[0150] 判断第一纠错结果是否为空，若是，则提示用户所输入的词语字符串所对应的汉字词语数目超过一个或者用户所输入的词语字符串有误，否则将纠错结果中的词语按照在词频表中的词频从大到小的排列输出设定个数的词语以提示给用户。

[0151] 在步骤S207中，根据第二纠错结果进行相应的提示，包括：

[0152] 1)若第二纠错结果为空，则提示用户所输入的词语字符串所对应的汉字词语数目超过一个或者用户所输入的词语字符串中的拼音有误；

[0153] 2)若第二纠错结果中词语的个数为1，则将第二纠错结果中的词语输出以提示给用户；

[0154] 3)若第二纠错结果中词语的个数大于1，则根据第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度确定将第二纠错结果中的各词语向用户进行提示的方式，并进行提示。

[0155] 进一步的，在上述的第3)种情况中，根据第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度确定将第二纠错结果中的各词语向用户进行提示的方式，并进行提示，包括：

[0156] 分别计算第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度；

[0157] 若计算出的第二纠错结果中各词语对应的所述相似度数值均一致，则将第二纠错结果中的各词语按照在词频表中的词频从大到小的排列输出设定个数的词语以提示给用户，否则将第二纠错结果中的各词语按照相似度从大到小输出设定个数的词语以提示给用户。

[0158] 在上述的第1)种和第3)种情况中，所述输入的词语字符串中的汉字词语字符串的获取过程包括：在分词处理的结果中分词个数大于1的情况下，依次记录所述输入的词语字符串的分词中的所有汉字分词并组成汉字词语字符串。

[0159] 下面以正确的原词为“火中取栗”为例，介绍一下本发明实施例的所述方法的应用过程：

[0160] 若用户输入存在近音别字且混有拼音的词语字符串“火中确1i”，因为该词语字符串中同时包含汉字和拼音，所以判定输入的词语字符串不是全拼音，然后对“火中确1i”进行分词处理，得到的分词为“火”“中”“确”“1i”，由于分词的个数大于1，将“火中确1i”转成

全拼音为“huozhongqueli”，再对“huozhongqueli”进行拼音纠错处理，具体是首选根据拼音反查表判断该全拼音拼写是否正确，由于在拼音反查表中包含的都是正确的拼音，在拼音反查表中没有找到“huozhongqueli”，故该全拼音拼写错误，则根据“huozhongqueli”与拼音反查表中的每个拼音之间的相似度中最高者“huozhongquli”，得到与“huozhongquli”对应的词语“火中取栗”作为第二纠错结果。假设第二纠错结果中还包括“火种取栗”，接下来，用第二纠错结果中的“火中取栗”、“火种取栗”分别与由“火中确li”的分词中的所有汉字分词组成的汉字词语字符串“火中确”计算相似度，从而确定出“火中取栗”才是最终的提示内容，以询问用户。

[0161] 本发明实施例巧妙地将相似度计算方法应用于拼音字符的相似度计算和中文字符的相似度计算中，并采用了不同的评价指标，分别是在进行拼音字符的相似度计算时使用的相似度阈值、以及在向用户提示纠错结果时用到相似度从大到小排列后按照设定个数进行输出，以得到比较准确的计算结果和输出结果。

[0162] 本发明实施例的所述输入纠错方法中，通过拼音纠错处理和中文字符的相似度计算相结合，比第一、二实施例更加准确、有效的处理同音别字、近音别字、拼音形近别字、多字漏字等词语输入错误问题，将本发明实施例的所述方法应用于中文搜索引擎和智能问答系统中，可以显著提高中文搜索引擎和智能问答系统中针对词语输入的信息查询和问答的准确率。

[0163] 本发明第四实施例，与第一实施例对应，本实施例介绍一种输入纠错装置，如图4所示，包括以下组成部分：

[0164] 1)判断模块401，用于判断输入的词语字符串是否为全拼音；若是，则将所述词语字符串的全拼音发送给纠错模块404进行处理，得到第一纠错结果；否则，将所述词语字符串发送给分词模块402进行分词处理；

[0165] 2)分词模块402，用于对判断模块401发来的词语字符串进行分词处理，在分词处理的结果中分词个数大于1时，将所述词语字符串发送给转换模块403；若分词处理的结果中分词的个数为1，则将用户输入的词语字符串作为第二纠错结果。

[0166] 3)转换模块403，用于将分词模块402发来的词语字符串转成全拼音并将转成的全拼音发送给纠错模块404进行处理，得到第二纠错结果；

[0167] 具体的，转换模块403用于：

[0168] 在不改变所述词语字符串中各分词出现顺序的情况下，将所述词语字符串中的汉字分词转换成拼音，再与所述词语字符串中已有的拼音一起，组成所述词语字符串对应的全拼音。

[0169] 4)纠错模块404，用于对判断模块401或者转换模块403发来的全拼音进行拼音纠错处理，并将纠错后的全拼音反查为汉字，得到相应的纠错结果。

[0170] 本发明实施例的所述输入纠错装置，通过拼音纠错处理，可有效处理出现同音别字、近音别字、拼音、形近别字、多字漏字等词语输入错误问题，将本发明实施例的所述装置应用于中文搜索引擎和智能问答系统中，可以显著提高中文搜索引擎和智能问答系统中针对词语输入的信息查询和问答的准确率。

[0171] 本发明第五实施例，与第二实施例对应，本实施例介绍一种输入纠错装置，如图5所示，包括以下组成部分：

[0172] 1)建立模块501,用于:提供训练语料;对训练语料进行分词以得到词语列表;在词语列表的基础上利用拼音反查表生成工具生成拼音反查表,采用统计的方式根据词语列表得到词频表。

[0173] 2)判断模块502,用于判断输入的词语字符串是否为全拼音;若是,则将所述词语字符串的全拼音发送给纠错模块505进行处理,得到第一纠错结果;否则,将所述词语字符串发送给分词模块503进行分词处理;

[0174] 3)分词模块503,用于对判断模块502发来的词语字符串进行分词处理,在分词处理的结果中分词个数大于1时,将所述词语字符串发送给转换模块504;若分词处理的结果中分词的个数为1,则将用户输入的词语字符串作为第二纠错结果。

[0175] 4)转换模块504,用于将分词模块503发来的词语字符串转成全拼音并将转成的全拼音发送给纠错模块505进行处理,得到第二纠错结果;

[0176] 具体的,转换模块504用于:

[0177] 在不改变所述词语字符串中各分词出现顺序的情况下,将所述词语字符串中的汉字分词转换成拼音,再与所述词语字符串中已有的拼音一起,组成所述词语字符串对应的全拼音。

[0178] 5)纠错模块505,用于对判断模块502或者转换模块504发来的全拼音进行拼音纠错处理,并根据拼音反查表将纠错后的全拼音反查为汉字,得到相应的纠错结果。

[0179] 具体的,纠错模块505,包括:

[0180] 拼写检查单元51,用于对任一词语字符串的全拼音进行拼音纠错处理时,根据拼音反查表判断所述任一词语字符串的全拼音的拼写是否正确;若是,则调用第一处理单元51,否则调用第二处理单元52;具体是用所述任一词语字符串的全拼音在拼音反查表中进行比对查找,若有一致的,则表明拼写正确,否则拼写错误。

[0181] 第一处理单元52,用于根据所述任一词语字符串的全拼音获取同音的词语列表,基于获取的同音的词语列表得到纠错结果;获取同音的词语列表会用到建立模块501所建立的词语列表和拼音反查表。

[0182] 第二处理单元53,用于根据所述任一词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度,确定出纠错结果。

[0183] 进一步的,第一处理单元52用于:

[0184] 判断获取的同音的词语列表是否为空,若是,则得到的纠错结果为空,否则将获取的同音的词语列表中的词语作为纠错结果;

[0185] 第二处理单元53用于:

[0186] 依次计算所述词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度;对于计算出的相似度大于设定相似度阈值的情况,将拼音反查表中的拼音对应的词语列表中的词语作为纠错结果。

[0187] 本发明实施例巧妙地将相似度计算应用于拼音字符的相似度计算和中文字符的相似度计算中,并采用了不同的评价指标,分别是在进行拼音字符的相似度计算时使用的相似度阈值、以及在向用户提示纠错结果时用到基于中文相似度从大到小排列纠错结果中的词语进行输出,以得到比较准确的计算结果和输出结果。

[0188] 本发明实施例的所述输入纠错装置,通过拼音纠错处理,可有效处理出现同音别

字、近音别字、拼音、形近别字、多字漏字等词语输入错误问题，将本发明实施例的所述装置应用于中文搜索引擎和智能问答系统中，可以显著提高中文搜索引擎和智能问答系统中针对词语输入的信息查询和问答的准确率。

[0189] 本发明第六实施例，与第三实施例对应，本实施例介绍一种输入纠错装置，如图6所示，包括以下组成部分：

[0190] 1)建立模块501，用于：提供训练语料；对训练语料进行分词以得到词语列表；在词语列表的基础上利用拼音反查表生成工具生成拼音反查表，采用统计的方式根据词语列表得到词频表。

[0191] 2)判断模块502，用于判断输入的词语字符串是否为全拼音；若是，则将所述词语字符串的全拼音发送给纠错模块505进行处理，得到第一纠错结果；否则，将所述词语字符串发送给分词模块503进行分词处理；

[0192] 3)分词模块503，用于对判断模块502发来的词语字符串进行分词处理，在分词处理的结果中分词个数大于1时，将所述词语字符串发送给转换模块504；若分词处理的结果中分词的个数为1，则将用户输入的词语字符串作为第二纠错结果。

[0193] 4)转换模块504，用于将分词模块503发来的词语字符串转成全拼音并将转成的全拼音发送给纠错模块505进行处理，得到第二纠错结果；

[0194] 具体的，转换模块504用于：

[0195] 在不改变所述词语字符串中各分词出现顺序的情况下，将所述词语字符串中的汉字分词转换成拼音，再与所述词语字符串中已有的拼音一起，组成所述词语字符串对应的全拼音。

[0196] 5)纠错模块505，用于对判断模块502或者转换模块504发来的全拼音进行拼音纠错处理，并根据拼音反查表将纠错后的全拼音反查为汉字，得到相应的纠错结果。

[0197] 具体的，纠错模块505，包括：

[0198] 拼写检查单元51，用于对任一词语字符串的全拼音进行拼音纠错处理时，根据拼音反查表判断所述任一词语字符串的全拼音的拼写是否正确；若是，则调用第一处理单元51，否则调用第二处理单元52；具体是用所述任一词语字符串的全拼音在拼音反查表中进行比对查找，若有一致的，则表明拼写正确，否则拼写错误。

[0199] 第一处理单元52，用于根据所述任一词语字符串的全拼音获取同音的词语列表，基于获取的同音的词语列表得到纠错结果；获取同音的词语列表会用到建立模块501所建立的词语列表和拼音反查表。

[0200] 第二处理单元53，用于根据所述任一词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度，确定出纠错结果。

[0201] 进一步的，第一处理单元52用于：

[0202] 判断获取的同音的词语列表是否为空，若是，则得到的纠错结果为空，否则将获取的同音的词语列表中的词语作为纠错结果；

[0203] 第二处理单元53用于：

[0204] 依次计算所述词语字符串的全拼音与拼音反查表中的每个拼音之间的相似度；对于计算出的相似度大于设定相似度阈值的情况，将拼音反查表中的拼音对应的词语列表中的词语作为纠错结果。

[0205] 6)提示模块506,用于根据第一纠错结果或者第二纠错结果进行相应的提示。

[0206] 具体的,提示模块506一方面用于:

[0207] 判断第一纠错结果是否为空,若是,则提示用户所输入的词语字符串所对应的汉字词语数目超过一个或者用户所输入的词语字符串有误,否则将纠错结果中的词语按照在词频表中的词频从大到小的排列输出设定个数的词语以提示给用户。

[0208] 提示模块506另一方面用于:

[0209] 若第二纠错结果为空,则提示用户所输入的词语字符串所对应的汉字词语数目超过一个或者用户所输入的词语字符串中的拼音有误;

[0210] 若第二纠错结果中词语的个数为1,则将第二纠错结果中的词语输出以提示给用户;

[0211] 若第二纠错结果中词语的个数大于1,则根据第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度确定将第二纠错结果中的各词语向用户进行提示的方式,并进行提示。

[0212] 进一步的,提示模块506在根据第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度确定将第二纠错结果中的各词语向用户进行提示的方式时,用于:

[0213] 分别计算第二纠错结果中各词语与所述输入的词语字符串中的汉字词语字符串之间的相似度;

[0214] 若计算出的第二纠错结果中各词语对应的所述相似度数值均一致,则将第二纠错结果中的各词语按照在词频表中的词频从大到小的排列输出设定个数的词语以提示给用户,否则将第二纠错结果中的各词语按照相似度从大到小输出设定个数的词语以提示给用户。

[0215] 在本发明实施例中,分词模块503,还用于:在分词处理的结果中分词个数大于1的情况下,依次记录所述输入的词语字符串的分词中的所有汉字分词并组成汉字词语字符串,发送给提示模块506。

[0216] 本发明实施例巧妙地将相似度计算方法应用于拼音字符的相似度计算和中文字符的相似度计算中,并采用了不同的评价指标,分别是在进行拼音字符的相似度计算时使用的相似度阈值、以及在向用户提示纠错结果时用到相似度从大到小排列后按照设定个数进行输出,以得到比较准确的计算结果和输出结果。

[0217] 本发明实施例的所述输入纠错装置,通过拼音纠错处理和中文字符的相似度计算相结合,比第四、五实施例更加准确、有效的处理同音别字、近音别字、拼音形近别字、多字漏字等词语输入错误问题,将本发明实施例的所述装置应用于中文搜索引擎和智能问答系统中,可以显著提高中文搜索引擎和智能问答系统中针对词语输入的信息查询和问答的准确率。

[0218] 本发明第七实施例,本实施例是在上述实施例的基础上,结合附图7～8介绍一个本发明的应用实例。

[0219] 本发明实施例提供一种基于分词和相似度计算的输入纠错方法及系统,可以用于解决用户在输入词语时可能会发生的因为同音别字、近音别字、形近别字、拼音、多字漏字等导致的词语输入错误问题,从而提高用户查询的准确率,使用户得到所需要的信息。

[0220] 如图7所示，为实现上述的目的，本发明实施例提供了一种基于分词和相似度计算的输入纠错方法，包括：

[0221] 步骤1：输入待查的词语字符串。

[0222] 步骤2：判断词语纠错模型是否已构建，如果是则进入步骤4，否则进入步骤3。

[0223] 步骤3：根据训练语料构建词语纠错模型。

[0224] 步骤4：根据词语纠错模型获取词语列表、词频表和拼音反查表。该拼音反查表是指根据拼音反查中文词语的列表。

[0225] 步骤5：判断所输入的词语字符串是否为全拼音，如果是则进入步骤6，否则进入步骤9。

[0226] 步骤6：对所输入的词语字符串进行拼音纠错处理，得到词语纠错结果列表；

[0227] 步骤7：判断词语纠错结果列表是否为空，如果是，则进入步骤8，否则进入步骤18；

[0228] 步骤8：提示用户所输入的拼音所对应的汉字词语数目超过一个或者用户所输入的拼音有误，进入步骤20；

[0229] 步骤9：对所输入的词语字符串进行分词处理；

[0230] 步骤10：判断分词处理结果中分词的个数是否等于1，如果是则进入步骤11，否则表明分词的个数大于1，进入步骤12；

[0231] 步骤11：将分词处理结果中的分词输出，并提示用户所输入的词语没有错误，进入步骤20；

[0232] 步骤12：按顺序记录所输入的词语字符串中出现的中文字符，将所输入的词语字符串转换为全拼音；

[0233] 步骤13：对转换后的全拼音字符串进行拼音纠错处理，得到词语纠错结果列表；

[0234] 步骤14：判断词语纠错结果列表是否为空，如果是则进入步骤8，否则进入步骤15；

[0235] 步骤15：判断词语纠错结果列表中的词语个数是否为1，如果是则进入步骤18，否则进入步骤16；

[0236] 步骤16：计算词语纠错结果列表中的词语和步骤12中所记录的中文字符之间的相似度；

[0237] 步骤17：判断词语纠错结果列表中的各词语对应的相似度值是否一样，如果是则进入步骤18，否则进入步骤19；

[0238] 步骤18：按照所得到的词语纠错结果列表中的词语在词频表中的词频从大到小进行结果输出并提示用户，进入步骤20；

[0239] 步骤19：提示并输出相似度值最大的词语，输入纠错流程结束；

[0240] 步骤20：输入纠错流程结束。

[0241] 优选地，所述步骤3中的词语纠错模型的构建，主要包括：

[0242] 步骤3.1：判断训练语料的分词结果文件是否存在，如果是则进入步骤3.3，否则进入步骤3.2；

[0243] 步骤3.2：对训练语料进行分词并将分词结果保存至分词结果文件；

[0244] 步骤3.3：基于分词结果文件统计词语列表和词频表；

[0245] 步骤3.4：判断拼音反查文件是否存在，如果是则进入步骤3.5，否则进入步骤3.6；

[0246] 步骤3.5：读取拼音反查表，进入步骤3.7；

[0247] 步骤3.6：在词语列表的基础上利用拼音反查工具获取拼音反查表，并保存至拼音反查文件中；

[0248] 步骤3.7：词语纠错模型构建完成。

[0249] 优选地，所述步骤6和步骤13中的拼音纠错处理，主要包括：

[0250] 步骤6.1：接收所输入的全拼音字符串；

[0251] 步骤6.2：判断拼音拼写是否正确，如果是则进入步骤6.3，否则进入步骤6.10；

[0252] 步骤6.3：根据拼音获取同音词语列表；

[0253] 步骤6.4：判断获取的同音词语列表是否为空，如果是则进入步骤6.8，否则进入步骤6.6；

[0254] 步骤6.5：输出空的同音词语列表作为词语纠错结果列表，进入步骤6.10；

[0255] 步骤6.6：根据词频表按词频从大到小重新排序所获取的词语列表；

[0256] 步骤6.7：输出获取的词语列表中设定个数的词语作为词语纠错结果列表，进入步骤6.10；

[0257] 步骤6.8：计算所输入的拼音分别与拼音反查表中的拼音之间的相似度，并按相似度大小排序；

[0258] 步骤6.9：获取相似度大于设定阈值的近音词语列表，输出获取的近音词语列表中设定个数的词语作为词语纠错结果列表；

[0259] 步骤6.10：拼音纠错处理结束。

[0260] 为实现上述的目的，本发明还提供了一种基于分词和相似度计算的中文词语纠错系统，采用下述的技术方案。

[0261] 如图8所示，一种基于分词和相似度计算的中文词语纠错系统，包括：文本输入模块、中文分词模块、拼音转化模块、词语纠错模型构建模块、相似度计算模块、拼音纠错处理模块、纠错结果筛选模块和用户提示模块；

[0262] 所述的文本输入模块，提供一个文本输入框，用于接收用户输入词语字符串；

[0263] 所述的中文分词模块，用于对训练语料进行分词以获取词语列表，以及用于对文本输入模块中所输入的词语字符串进行分词；

[0264] 所述的拼音转化模块，用于提供拼音转换功能，对词语列表进行拼音转换以得到拼音反查表，以及对文本输入模块中所输入的词语进行拼音转换从而得到拼音字符串；

[0265] 所述的词语纠错模型构建模块，用于构建词语纠错模型，接收中文分词模块对训练语料处理后的词语列表和统计相应的词频表，并将词语列表保存至分词结果文件中，以及接收拼音转化模块所得到的拼音反查表并保存至拼音反查文件中；

[0266] 本发明实施例中提供的词语纠错模型构建模块，可以有效充分利用训练语料所提供的词语信息，并从中获取到词语列表、词频表和拼音反查表，可快速适应不同领域的自定义词语纠错模型的构建。

[0267] 所述的相似度计算模块，采用了基于编辑距离的相似度计算方法，用于对拼音字符串和词语纠错模型构建模块中所得到的拼音反查表中的拼音进行相似度计算并取其中大于设定阈值的词语集合，以及对拼音纠错处理模块所得到的词语纠错结果列表中各词语与所述用户输入词语字符串中的汉字词语字符串之间的相似度并取其中相似度值最大的一个或几个词语；

[0268] 所述的拼音纠错处理模块，用于对拼音转换模块所得到的拼音字符串进行纠错处理，包括同音词语处理、近音词语处理和拼音处理，从而获得词语纠错结果列表；

[0269] 所述的纠错结果筛选模块，用于对拼音纠错处理模块所得到的词语纠错结果列表进行筛选、排序等处理，可分别按照相似度大小和词频大小进行排序输出供用户提示模块所使用的词语纠错结果列表；

[0270] 所述的用户提示模块，用于对中文词语纠错结果进行输出并提示用户，包括不存在错误时提示用户输入的词语没有错误、词语纠错结果列表为空时提示用户输入拼音有误或不是输入的词语不只一个，按照纠错结果筛选模块处理后的词语纠错结果列表输出。

[0271] 通过具体实施方式的说明，应当可对本发明为达成预定目的所采取的技术手段及功效得以更加深入且具体的了解，然而所附图示仅是提供参考与说明之用，并非用来对本发明加以限制。

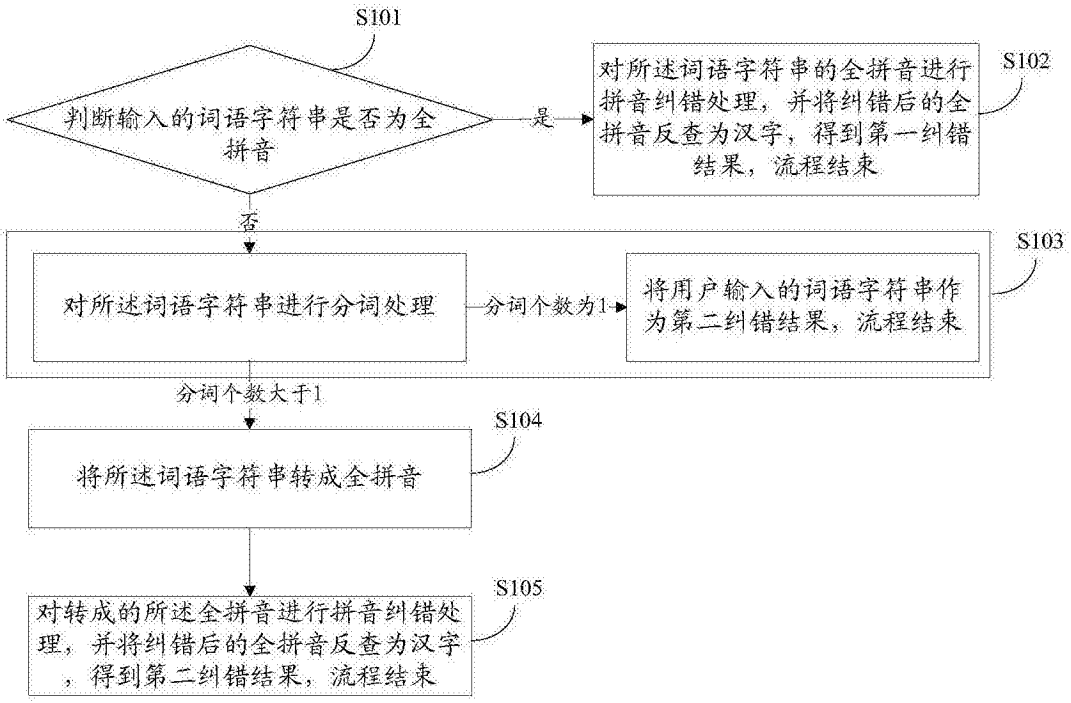

S101
判断输入的词语字符串是否为全拼音
是
对所述词语字符串的全拼音进行拼音纠错处理，并将纠错后的全拼音反查为汉字，得到第一纠错结果，流程结束
S102
否
对所述词语字符串进行分词处理
分词个数为1
将用户输入的词语字符串作为第二纠错结果，流程结束
S103
分词个数大于1
S104
将所述词语字符串转成全拼音
S105
对转成的所述全拼音进行拼音纠错处理，并将纠错后的全拼音反查为汉字，得到第二纠错结果，流程结束

图1

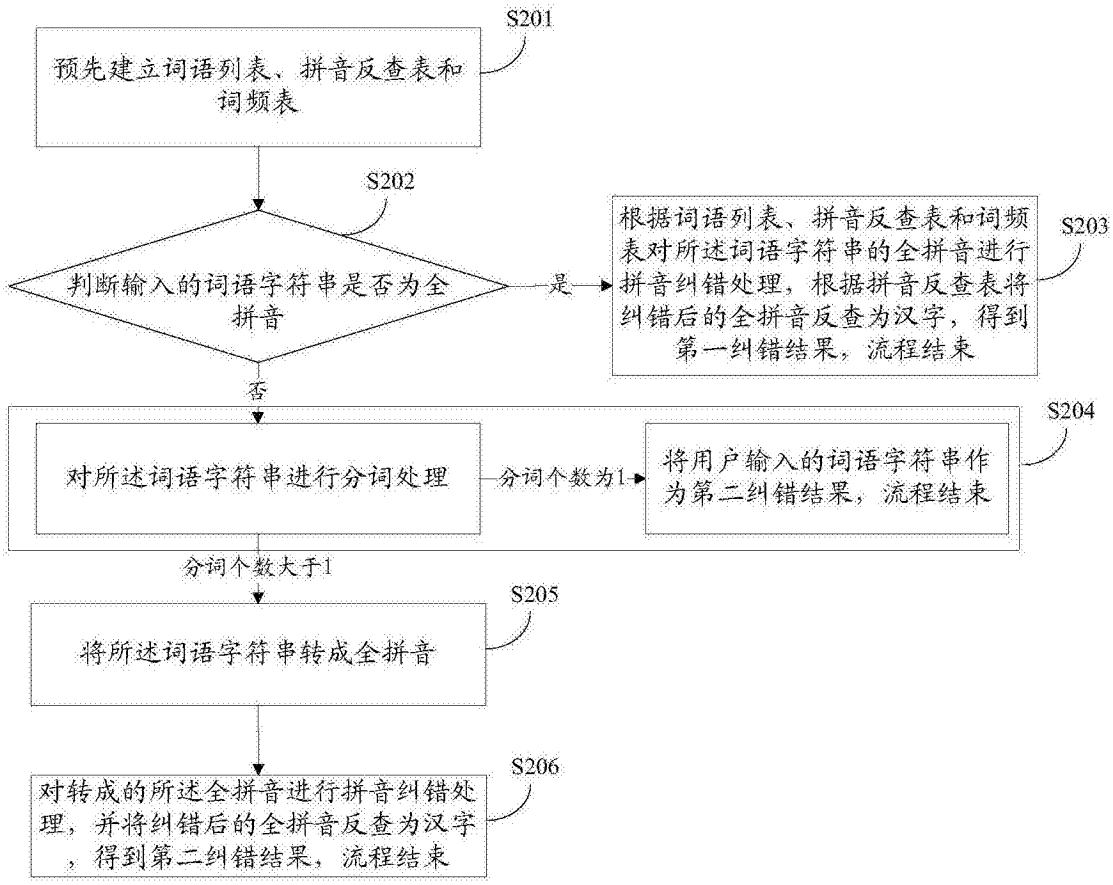

S201
预先建立词语列表、拼音反查表和词频表
S202
判断输入的词语字符串是否为全拼音
是
根据词语列表、拼音反查表和词频表对所述词语字符串的全拼音进行拼音纠错处理，根据拼音反查表将纠错后的全拼音反查为汉字，得到第一纠错结果，流程结束
S203
否
对所述词语字符串进行分词处理
分词个数为1
将用户输入的词语字符串作为第二纠错结果，流程结束
S204
分词个数大于1
S205
将所述词语字符串转成全拼音
S206
对转成的所述全拼音进行拼音纠错处理，并将纠错后的全拼音反查为汉字，得到第二纠错结果，流程结束

图2

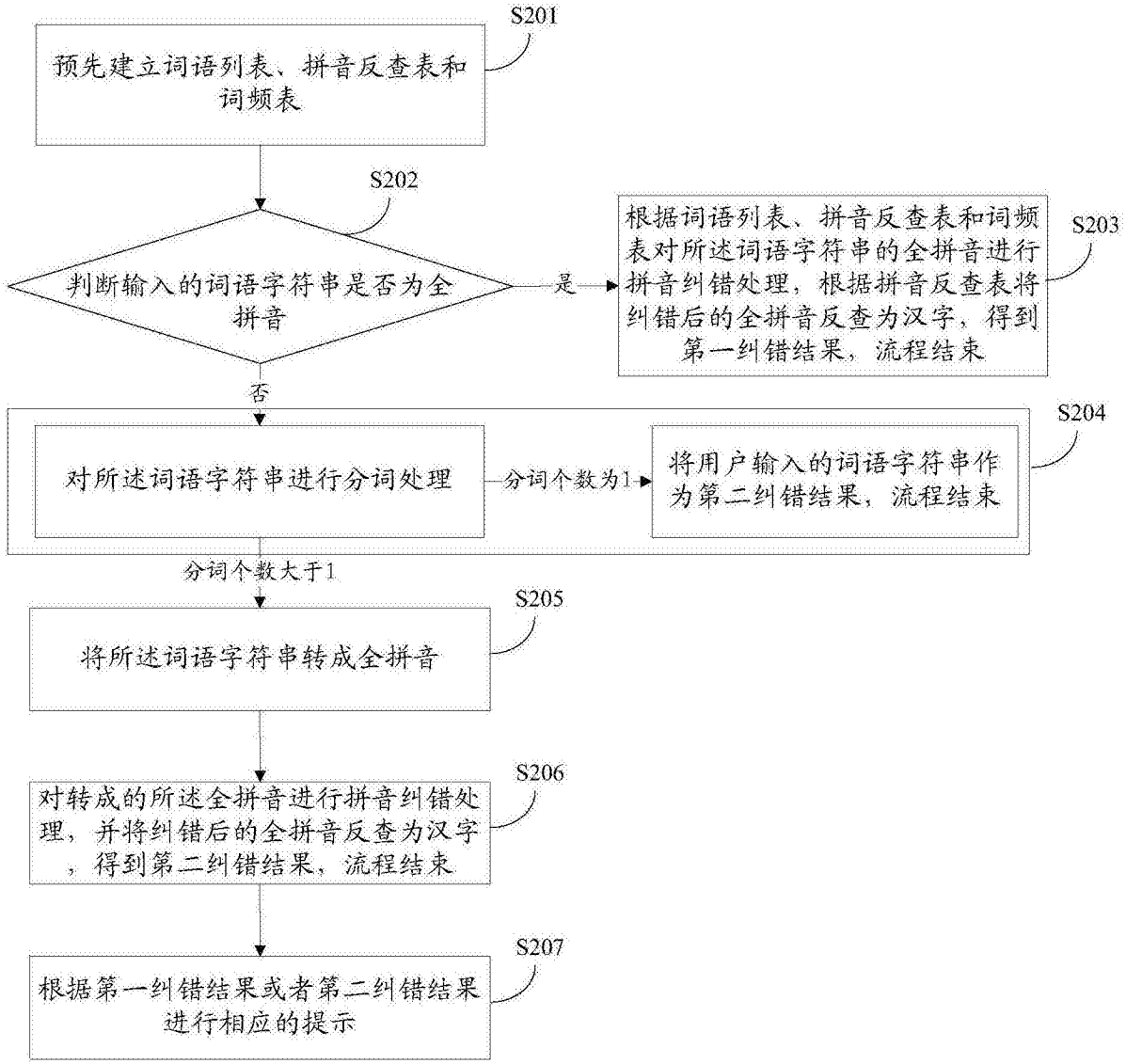

S201
预先建立词语列表、拼音反查表和词频表
S202
判断输入的词语字符串是否为全拼音
是
S203
根据词语列表、拼音反查表和词频表对所述词语字符串的全拼音进行拼音纠错处理，根据拼音反查表将纠错后的全拼音反查为汉字，得到第一纠错结果，流程结束
否
S204
对所述词语字符串进行分词处理
分词个数为1
将用户输入的词语字符串作为第二纠错结果，流程结束
分词个数大于1
S205
将所述词语字符串转成全拼音
S206
对转成的所述全拼音进行拼音纠错处理，并将纠错后的全拼音反查为汉字，得到第二纠错结果，流程结束
S207
根据第一纠错结果或者第二纠错结果进行相应的提示

图3

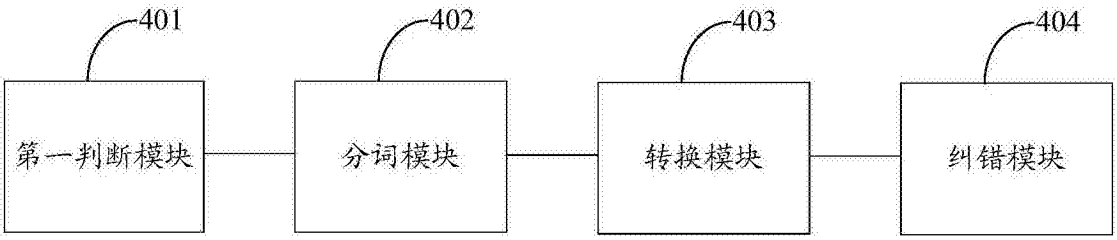

401
第一判断模块
402
分词模块
403
转换模块
404
纠错模块

图4

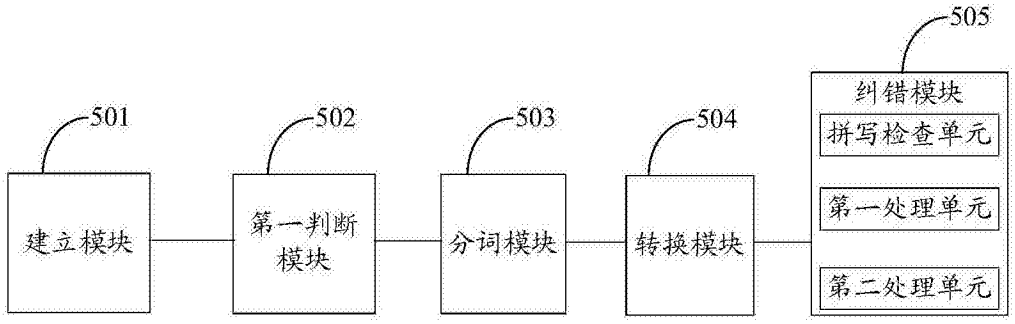

501
建立模块
502
第一判断模块
503
分词模块
504
转换模块
505
纠错模块
拼写检查单元
第一处理单元
第二处理单元

图5

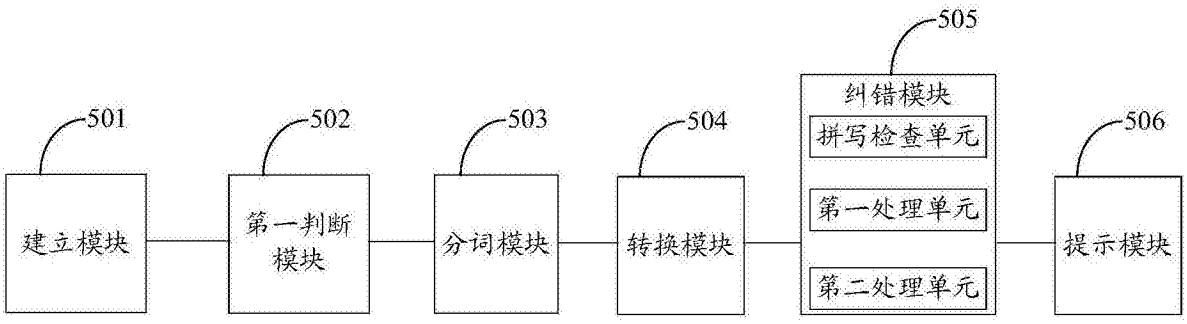

501
建立模块
502
第一判断模块
503
分词模块
504
转换模块
505
纠错模块
拼写检查单元
第一处理单元
第二处理单元
506
提示模块

图6

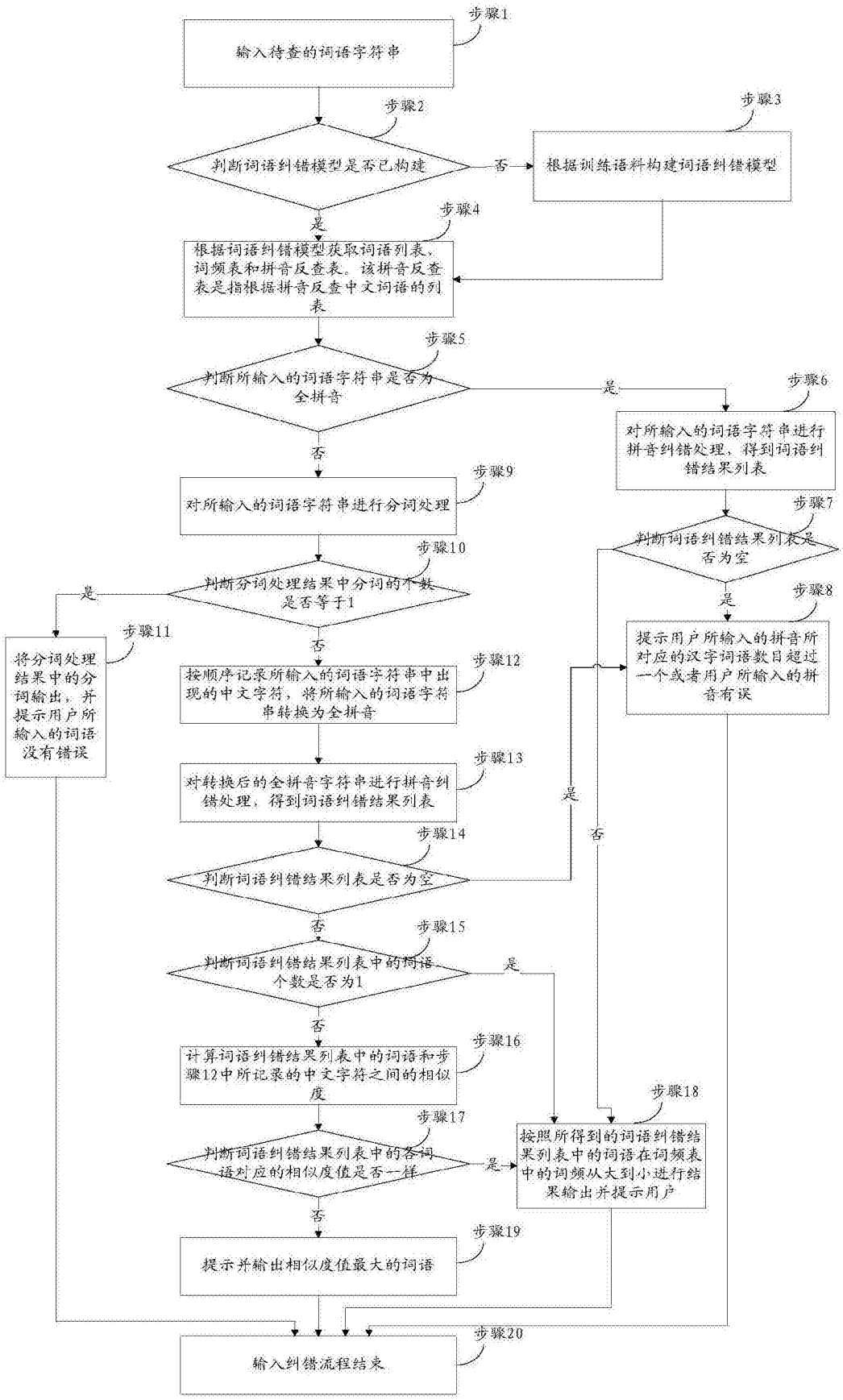

步骤1
输入待查的词语字符串
步骤2
判断词语纠错模型是否已构建
否
步骤3
根据训练语料构建词语纠错模型
是
步骤4
根据词语纠错模型获取词语列表、词频表和拼音反查表。该拼音反查表是指根据拼音反查中文词语的列表
步骤5
判断所输入的词语字符串是否为全拼音
是
步骤6
对所输入的词语字符串进行拼音纠错处理，得到词语纠错结果列表
否
步骤9
对所输入的词语字符串进行分词处理
步骤7
判断词语纠错结果列表是否为空
步骤10
判断分词处理结果中分词的个数是否等于1
是
步骤8
提示用户所输入的拼音所对应的汉字词语数目超过一个或者用户所输入的拼音有误
步骤11
将分词处理结果中的分词输出，并提示用户所输入的词语没有错误
否
步骤12
按顺序记录所输入的词语字符串中出现的中文字符，将所输入的词语字符串转换为全拼音
步骤13
对转换后的全拼音字符串进行拼音纠错处理，得到词语纠错结果列表
步骤14
判断词语纠错结果列表是否为空
是
否
步骤15
判断词语纠错结果列表中的词语个数是否为1
是
否
步骤16
计算词语纠错结果列表中的词语和步骤12中所记录的中文字符之间的相似度
步骤18
步骤17
判断词语纠错结果列表中的各词语对应的相似度值是否一样
是
按照所得到的词语纠错结果列表中的词语在词频表中的词频从大到小进行结果输出并提示用户
否
步骤19
提示并输出相似度值最大的词语
步骤20
输入纠错流程结束

图7

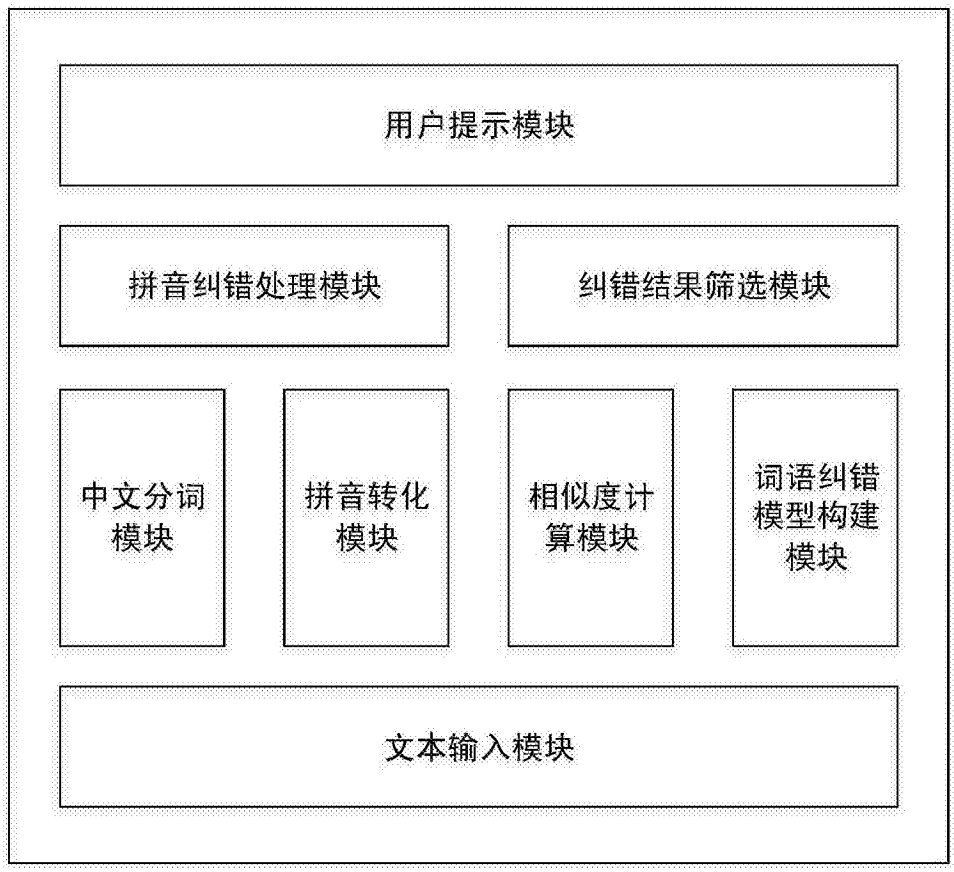

用户提示模块
拼音纠错处理模块
纠错结果筛选模块
中文分词模块
拼音转化模块
相似度计算模块
词语纠错模型构建模块
文本输入模块

图8