(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 106326484 A
(43)申请公布日 2017.01.11

(21)申请号 201610799830.4

(22)申请日 2016.08.31

(71)申请人 北京奇艺世纪科技有限公司
地址 100080 北京市海淀区北一街2号鸿城拓展大厦10、11层

(72)发明人 胡军 陈英傑 王天畅 叶澄灿

(74)专利代理机构 北京润泽恒知识产权代理有限公司 11319
代理人 苏培华

(51)Int.Cl.
G06F 17/30(2006.01)
G06F 17/27(2006.01)

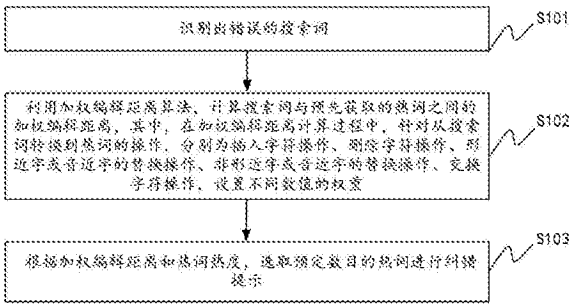
权利要求书3页 说明书10页 附图2页

(54)发明名称

搜索词纠错方法及装置

(57)摘要

本发明提供了一种搜索词纠错方法及装置，其中的方法包括：识别出错误的搜索词；利用加权编辑距离算法，计算所述搜索词与预先获取的热词之间的加权编辑距离，其中，在所述加权编辑距离计算过程中，针对从搜索词转换到热词的操作，分别为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作，设置不同数值的权重；根据所述加权编辑距离和热词热度，选取预定数目的热词进行纠错提示。本发明可提高对错误搜索词的纠错准确率。



CN 106326484 A

1.一种搜索词纠错方法，其特征在于，包括：

识别出错误的搜索词；

利用加权编辑距离算法，计算所述搜索词与预先获取的热词之间的加权编辑距离，其中，在所述加权编辑距离计算过程中，针对从搜索词转换到热词的操作，分别为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作，设置不同数值的权重；

根据所述加权编辑距离和热词热度，选取预定数目的热词进行纠错提示。

2.根据权利要求1所述的方法，其特征在于，所述利用加权编辑距离算法，计算所述搜索词与预先获取的热词之间的加权编辑距离，包括：

定义状态转移方程，用于表示所述搜索词与热词之间的加权编辑距离，其中，在状态转移方程中定义两个状态量，用于分别表示搜索词和热词之间对应位置的字符；

根据为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作所设置的不同数值的权重，求解所述状态转移方程在相应操作的解；

根据所述状态转移方程的解，得到所述加权编辑距离。

3.根据权利要求2所述的方法，其特征在于，所述状态转移方程为：

$$edit(i,j) = \min\{edit(i-1,j)+1, edit(i,j-1)+1, edit(i-1,j-1)+f(i,j),\};$$

其中，i、j为所述两个状态量，f(i,j)为操作代价值，f(i,j)根据为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作所设置的不同数值的权重，得到各权重对应的代价值。

4.根据权利要求1-3任一项所述的方法，其特征在于，还包括：

通过查找预先设置的形近字映射表或音近字映射表，确定所述搜索词与所述热词是否互为形近字或音近字。

5.根据权利要求1-3任一项所述的方法，其特征在于，还包括：

设置各操作权重满足如下关系：

形近字或音近字的替换操作权重<交换字符操作权重<插入字符操作权重＝删除字符操作权重＝非形近字或音近字的替换操作权重。

6.根据权利要求1-3任一项所述的方法，其特征在于，所述识别出错误的搜索词，包括：

基于搜索日志，解析或计算出待识别搜索词的搜索点击率、词特征、出现概率、全匹配结果数和全匹配占比；

根据待识别搜索词的所述搜索点击率、所述词特征、所述出现概率、所述全匹配结果数和所述全匹配占比，确定所述待识别搜索词为错误搜索词或正常搜索词。

7.根据权利要求1-3任一项所述的方法，其特征在于，所述根据所述加权编辑距离和热词热度，选取预定数目的热词进行纠错提示，包括：

将热词搜索次数进行归一化处理；

根据所述加权编辑距离与热词搜索次数归一化处理结果，计算推荐综合得分；

选择推荐综合得分最高且所述加权编辑距离小于预定值的预定数目的热词，作为纠错

的推荐词，进行纠错提示。

8.一种加权编辑距离计算方法，其特征在于，包括：

获取源字符串和目标字符串；

计算所述源字符串和所述目标字符串之间的加权编辑距离，其中，在所述加权编辑距离计算过程中，针对从所述源字符串转换到所述目标字符串的不同操作分别设置不同数值的权重。

9.根据权利要求8所述的方法，其特征在于，所述计算所述源字符串和所述目标字符串之间的加权编辑距离，包括：

定义状态转移方程，用于表示所述源字符串和所述目标字符串之间的加权编辑距离，其中，在状态转移方程中定义两个状态量，用于分别表示源字符串和所述目标字符串之间对应位置的字符；

根据为不同操作所设置的不同数值的权重，求解所述状态转移方程在相应操作的解；

根据所述状态转移方程的解，得到所述加权编辑距离。

10.一种搜索词纠错装置，其特征在于，包括：

错误搜索词识别单元，用于识别出错误的搜索词；

加权编辑距离计算单元，用于利用加权编辑距离算法，计算所述搜索词与预先获取的热词之间的加权编辑距离，其中，在所述加权编辑距离计算过程中，针对从搜索词转换到热词的操作，分别为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作，设置不同数值的权重；

纠错提示单元，用于根据所述加权编辑距离和热词热度，选取预定数目的热词进行纠错提示。

11.根据权利要求10所述的装置，其特征在于，所述加权编辑距离计算单元包括：

状态转移方程定义子单元，用于定义状态转移方程，用于表示所述搜索词与热词之间的加权编辑距离，其中，在状态转移方程中定义两个状态量，用于分别表示搜索词和热词之间对应位置的字符；

方程求解子单元，用于根据为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作所设置的不同数值的权重，求解所述状态转移方程在相应操作的解，作为所述加权编辑距离。

12.根据权利要求11所述的装置，其特征在于，所述状态转移方程为：

$$edit(i,j)=\min\{edit(i-1,j)+1,edit(i,j-1)+1,edit(i-1,j-1)+f(i,j),\};$$

其中，i、j为所述两个状态量，f(i,j)为操作代价值，f(i,j)根据为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作所设置的不同数值的权重，得到各权重对应的代价值。

13.根据权利要求10-12任一项所述的装置，其特征在于，还包括：

形近字或音近字确定单元，用于查找预先设置的形近字映射表或音近字映射表，确定所述搜索词与所述热词是否互为形近字或音近字。

14.根据权利要求10-12任一项所述的装置，其特征在于，还包括：

操作权重设置单元，用于设置各操作权重满足如下关系：

形近字或音近字的替换操作权重<交换字符操作权重<插入字符操作权重＝删除字符操作权重＝非形近字或音近字的替换操作权重。

15.根据权利要求10-12任一项所述的装置，其特征在于，所述错误搜索词识别单元包括：

日志查找及计算子单元，用于基于搜索日志，解析或计算出待识别搜索词的搜索点击率、词特征、出现概率、全匹配结果数和全匹配占比；

识别结果确定子单元，用于根据待识别搜索词的所述搜索点击率、所述词特征、所述出现概率、所述全匹配结果数和所述全匹配占比，确定所述待识别搜索词为错误搜索词或正常搜索词。

16.根据权利要求10-12任一项所述的装置，其特征在于，所述纠错提示单元包括：

归一化处理子单元，用于将热词搜索次数进行归一化处理；

推荐综合得分计算子单元，用于根据所述加权编辑距离与热词搜索次数归一化处理结果，计算推荐综合得分；

推荐词确定子单元，用于选择推荐综合得分最高且所述加权编辑距离小于预定值的预定数目的热词，作为纠错的推荐词，进行纠错提示。

17.一种加权编辑距离计算装置，其特征在于，包括：

获取单元，用于获取源字符串和目标字符串；

计算单元，用于计算所述源字符串和所述目标字符串之间的加权编辑距离，其中，在所述加权编辑距离计算过程中，针对从所述源字符串转换到所述目标字符串的不同操作分别设置不同数值的权重。

18.根据权利要求17所述的装置，其特征在于，所述计算单元包括：

状态转移方程定义子单元，用于定义状态转移方程，用于表示所述源字符串和所述目标字符串之间的加权编辑距离，其中，在状态转移方程中定义两个状态量，用于分别表示源字符串和所述目标字符串之间对应位置的字符；

状态转移方程求解子单元，用于根据为不同操作所设置的不同数值的权重，求解所述状态转移方程在相应操作的解，作为所述加权编辑距离。

搜索词纠错方法及装置

技术领域

[0001] 本发明涉及计算机技术领域，特别是涉及一种搜索词纠错方法及装置。

背景技术

[0002] 用户输入搜索词中往往包含大量的错别字，导致搜索结果不能满足用户的需求。错误搜索词产生的原因比较复杂，主要包括：同音字选字错误、拼音拼写错误、字形输入错误等。为了解决上述问题，可以采用纠错提示的方式，在搜索页面提示用户输入的搜索词可能不准确，并根据其输入的搜索词，推荐相关可能的搜索词。传统纠错技术，大多采用编辑距离技术将原词与词典中的词条比较，然后，选择与原词编辑距离最小的K个词条。编辑操作包括：1)将一个字符替换成另一个字符，2)插入一个字符，3)删除一个字符。这种编辑操作并未考虑替换字符之间的关系，很多情况下，这种传统的编辑距离效果并不是很好。

发明内容

[0003] 为了提高搜索词纠错准确率，本发明实施例提供一种搜索词纠错方法及装置。

[0004] 根据本发明一个方面，提供一种搜索词纠错方法，包括：识别出错误的搜索词；利用加权编辑距离算法，计算所述搜索词与预先获取的热词之间的加权编辑距离，其中，在所述加权编辑距离计算过程中，针对从搜索词转换到热词的操作，分别为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作，设置不同数值的权重；根据所述加权编辑距离和热词热度，选取预定数目的热词进行纠错提示。

[0005] 优选的，所述利用加权编辑距离算法，计算所述搜索词与预先获取的热词之间的加权编辑距离，包括：定义状态转移方程，用于表示所述搜索词与热词之间的加权编辑距离，其中，在状态转移方程中定义两个状态量，用于分别表示搜索词和热词之间对应位置的字符；根据为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作所设置的不同数值的权重，求解所述状态转移方程在相应操作的解；根据所述状态转移方程的解，得到所述加权编辑距离。

[0006] 优选的，所述状态转移方程为：$edit(i,j)=\min\{edit(i-1,j)+1,edit(i,j-1)+1,edit(i-1,j-1)+f(i,j),\}$；其中，i、j为所述两个状态量，f(i,j)为操作代价值，f(i,j)根据为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作所设置的不同数值的权重，得到各权重对应的代价值。

[0007] 优选的，还包括：通过查找预先设置的形近字映射表或音近字映射表，确定所述搜索词与所述热词是否互为形近字或音近字。

[0008] 优选的，还包括：设置各操作权重满足如下关系：形近字或音近字的替换操作权重<交换字符操作权重<插入字符操作权重＝删除字符操作权重＝非形近字或音近字的替换操作权重。

[0009] 优选的，所述识别出错误的搜索词，包括：基于搜索日志，解析或计算出待识别搜索词的搜索点击率、词特征、出现概率、全匹配结果数和全匹配占比；根据待识别搜索词的所述搜索点击率、所述词特征、所述出现概率、所述全匹配结果数和所述全匹配占比，确定所述待识别搜索词为错误搜索词或正常搜索词。

[0010] 优选的，所述根据所述加权编辑距离和热词热度，选取预定数目的热词进行纠错提示，包括：将热词搜索次数进行归一化处理；根据所述加权编辑距离与热词搜索次数归一化处理结果，计算推荐综合得分；选择推荐综合得分最高且所述加权编辑距离小于预定值的预定数目的热词，作为纠错的推荐词，进行纠错提示。

[0011] 根据本发明的另一个方面，提供一种加权编辑距离计算方法，包括：获取源字符串和目标字符串；计算所述源字符串和所述目标字符串之间的加权编辑距离，其中，在所述加权编辑距离计算过程中，针对从所述源字符串转换到所述目标字符串的不同操作分别设置不同数值的权重。

[0012] 优选的，所述计算所述源字符串和所述目标字符串之间的加权编辑距离，包括：定义状态转移方程，用于表示所述源字符串和所述目标字符串之间的加权编辑距离，其中，在状态转移方程中定义两个状态量，用于分别表示源字符串和所述目标字符串之间对应位置的字符；根据为不同操作所设置的不同数值的权重，求解所述状态转移方程在相应操作的解；根据所述状态转移方程的解，得到所述加权编辑距离。

[0013] 根据本发明的又一个方面，提供一种搜索词纠错装置，包括：错误搜索词识别单元，用于识别出错误的搜索词；加权编辑距离计算单元，用于利用加权编辑距离算法，计算所述搜索词与预先获取的热词之间的加权编辑距离，其中，在所述加权编辑距离计算过程中，针对从搜索词转换到热词的操作，分别为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作，设置不同数值的权重；纠错提示单元，用于根据所述加权编辑距离和热词热度，选取预定数目的热词进行纠错提示。

[0014] 优选的，所述加权编辑距离计算单元包括：状态转移方程定义子单元，用于定义状态转移方程，用于表示所述搜索词与热词之间的加权编辑距离，其中，在状态转移方程中定义两个状态量，用于分别表示搜索词和热词之间对应位置的字符；方程求解子单元，用于根据为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作所设置的不同数值的权重，求解所述状态转移方程在相应操作的解，作为所述加权编辑距离。

[0015] 优选的，所述状态转移方程为：$edit(i,j)=\min\{edit(i-1,j)+1, edit(i,j-1)+1, edit(i-1,j-1)+f(i,j),\}$；其中，i、j为所述两个状态量，f(i,j)为操作代价值，f(i,j)根据为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作所设置的不同数值的权重，得到各权重对应的代价值。

[0016] 优选的，还包括：形近字或音近字确定单元，用于查找预先设置的形近字映射表或音近字映射表，确定所述搜索词与所述热词是否互为形近字或音近字。

[0017] 优选的，还包括：操作权重设置单元，用于设置各操作权重满足如下关系：形近字或音近字的替换操作权重<交换字符操作权重<插入字符操作权重＝删除字符操作权重＝

非形近字或音近字的替换操作权重。

[0018] 优选的，所述错误搜索词识别单元包括：日志查找及计算子单元，用于基于搜索日志，解析或计算出待识别搜索词的搜索点击率、词特征、出现概率、全匹配结果数和全匹配占比；识别结果确定子单元，用于根据待识别搜索词的所述搜索点击率、所述词特征、所述出现概率、所述全匹配结果数和所述全匹配占比，确定所述待识别搜索词为错误搜索词或正常搜索词。

[0019] 优选的，所述纠错提示单元包括：归一化处理子单元，用于将热词搜索次数进行归一化处理；推荐综合得分计算子单元，用于根据所述加权编辑距离与热词搜索次数归一化处理结果，计算推荐综合得分；推荐词确定子单元，用于选择推荐综合得分最高且所述加权编辑距离小于预定值的预定数目的热词，作为纠错的推荐词，进行纠错提示。

[0020] 根据本发明的再一个方面，提供一种加权编辑距离计算装置，包括：获取单元，用于获取源字符串和目标字符串；计算单元，用于计算所述源字符串和所述目标字符串之间的加权编辑距离，其中，在所述加权编辑距离计算过程中，针对从所述源字符串转换到所述目标字符串的不同操作分别设置不同数值的权重。

[0021] 优选的，所述计算单元包括：状态转移方程定义子单元，用于定义状态转移方程，用于表示所述源字符串和所述目标字符串之间的加权编辑距离，其中，在状态转移方程中定义两个状态量，用于分别表示源字符串和所述目标字符串之间对应位置的字符；状态转移方程求解子单元，用于根据为不同操作所设置的不同数值的权重，求解所述状态转移方程在相应操作的解，作为所述加权编辑距离。

[0022] 可见，本发明实施例提供一种基于加权编辑距离的搜索词纠错方法及装置，通过为删除字符操作、插入字符操作、形近字或音近字替换操作、非形近字或音近字替换操作以及交换字符操作，分别设置不同的权重，从而在加权编辑距离计算过程中，充分涵盖了从搜索词到热词转换过程中可能涉及的各种操作，从而可更加快速、准确的计算出从搜索词到热词之间的编辑距离，提高搜索词纠错准确性。

附图说明

[0023] 图1是本发明一个实施例提供的一种搜索词纠错方法流程图；

[0024] 图2是本发明一个实施例提供的一种加权编辑距离计算方法流程图；

[0025] 图3是本发明一个实施例提供的一种搜索词纠错装置结构示意图。

具体实施方式

[0026] 为使本发明的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本发明作进一步详细的说明。

[0027] 如前分析的，现有技术中基于编辑距离的纠错方案并未考虑替换字符之间的关系，比如形近字、音近字等，也未考虑字符串内邻近字符之间的交换操作，因此这种传统的编辑距离效果并不理想。

[0028] 本发明实施例提供一种基于加权编辑距离的搜索词纠错方法及装置，通过为删除字符操作、插入字符操作、形近字或音近字替换操作、非形近字或音近字替换操作以及交换字符操作，分别设置不同的权重，从而在加权编辑距离计算过程中，充分涵盖了从搜索词到

热词转换过程中可能涉及的各种操作，从而可更加快速、准确的计算出从搜索词到热词之间的编辑距离，提高搜索词纠错准确性。

[0029] 参见图1，为本发明实施例提供的一种搜索词纠错方法流程图，该方法包括：

[0030] S101：识别出错误的搜索词。

[0031] 对搜索词进行纠错，是针对错误的搜索词进行纠错，因此首先需要识别出错误的搜索词。搜索词之所以错误，包括很多种情况，例如，因同音字选字错误产生的搜索词、因拼音拼写错误产生的搜索词、因字形输入错误产生的搜索词，这就导致搜索结果不能满足用户的需求。在具体实现中，可以基于搜索日志，识别出错误的搜索词：通过搜索词点击率、搜索结果完全匹配的结果个数、以及基于语言模型的搜索词概率，可有效识别出错误的搜索词。

[0032] 本发明实施例中，提供以下包括步骤1-2识别错误的搜索词的方法：

[0033] 步骤1、基于搜索日志，解析或计算出待识别搜索词的搜索点击率、词特征、出现概率、全匹配结果数和全匹配占比。

[0034] 具体的，

[0035] 首先，计算待识别搜索词的搜索点击率。例如，从搜索日志中获取用户针对待识别搜索词的搜索次数和点击搜索结果次数；将点击搜索结果次数除以搜索次数，得到搜索点击率。

[0036] 其次，对待识别搜索词进行分词处理，得到多个词特征。

[0037] 继而，利用统计语言模型和各个词特征，计算待识别搜索词出现的概率。

[0038] 然后，计算待识别搜索词的全匹配结果数和相关结果数，其中，全匹配结果数为针对待识别搜索词的所有搜索结果中包含待识别搜索词的全部内容的结果的个数，相关结果数为针对待识别搜索词的所有搜索结果中包含待识别搜索词的部分内容的结果的个数。

[0039] 最后，计算全匹配结果数和相关结果数的比值，得到全匹配占比。

[0040] 步骤2、根据待识别搜索词的搜索点击率、词特征、出现概率、全匹配结果数和全匹配占比，确定待识别搜索词为错误搜索词或正常搜索词。

[0041] 通过融合待识别搜索词的多维度特征(即搜索点击率、各个所述词特征、待识别搜索词出现的概率、全匹配结果数和全匹配占比)，并基于多维度特征对待识别搜索词进行识别，降低了对待识别搜索词进行识别的难度，从而提高了对待识别搜索词的识别能力，有利于识别出待识别搜索词是否为错误搜索词。

[0042] S102：利用加权编辑距离算法，计算搜索词与预先获取的热词之间的加权编辑距离，其中，在加权编辑距离计算过程中，针对从搜索词转换到热词的操作，分别为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作，设置不同数值的权重。

[0043] 热词，是指网络或当下常用或流行的词，在实际操作中，可通过点击率确定众多热词组成热词表。在热词表的产生过程中，需要过滤掉点击率低、搜索结果少的词语，以保证热词的准确性。本发明实施例中，通过将错误的搜索词纠正为编辑距离相近的热词，从而为用户提供更好的体验。

[0044] 本领域技术人员理解，编辑距离(Edit Distance)，又称Levenshtein距离，是指两个字串之间，由一个转成另一个所需的最少编辑操作次数。传统的编辑操作包括将一个字

符替换成另一个字符，插入一个字符，删除一个字符。一般来说，编辑距离越小，两个字符串的相似度越大。

[0045] 传统的编辑距离操作包括：将一个字符替换成另一个字符，插入一个字符，删除一个字符，每一种操作对应的距离都是1。这些操作中并不包括字符串内临近字符之间的交换操作，也没有考虑音近字或形近字替换操作的特殊性。交换操作可以通过两次替换操作实现，比如CD->DD->DC，按照传统编辑距离算法，对应的距离是2。考虑到实际搜索过程中，用户将两个字符输入错位的概率非常高，交换操作的距离为2是不合理的。另外，因此，用户出现音近字或形近字导致的搜索词输入错误的概率也较高，如果对此特殊的替换操作没有特别处理，也设置距离为1，显然也是不合理的。

[0046] 因此，本发明实施例中提出了一种加权编辑距离方法，包括以下5种操作，并分别设置不同的权重。

[0047] 1)插入字符操作，权重为1；

[0048] 2)删除字符操作，权重为1；

[0049] 3)非形近字或者音近字的替换操作，权重为1；

[0050] 4)形近字或音近字的替换操作，权重为w1；

[0051] 5)交换字符操作，权重为w2。

[0052] 为了得到更好的效果，w1和w2的取值需特别注意，一般而言，需要满足w1<w2<1。经过实验得到，优选的，设置各操作权重满足如下关系：形近字或音近字的替换操作权重<交换字符操作权重<插入字符操作权重＝删除字符操作权重＝非形近字或音近字的替换操作权重。

[0053] 由于上述考虑了各种可能的操作，并且分别赋予不同的权重，因此采取这种加权编辑距离应用于纠错，计算字符串之间的相似度，精度更高。

[0054] 加权编辑距离(也称“动态规划”)可为不同操作设置不同的权重，解决上述问题。其思路是：通过描述出操作的状态，并且以一个状态转移方程进行求解。

[0055] 对于编辑距离问题，牵涉到源字符串str1(本实施例中的搜索词)和目标字符串str2(本实施例中的热词)，显然一个状态量是不能描述这种两元关系，因此，就使用了i,j两个量来描述一个状态。对于编辑距离的某个状态，从源字符串str1的1->i到目标字符串str2的1->j的最优编辑距离用edit[i,j]来表示，那么，目标就是得到一个状态转移方程，即怎样从ti<i、tj<j的这些子状态转移到i、j。在本发明实施例的加权编辑距离的操作中，包括插入字符操作、删除字符操作、音近字或形近字字符替换操作、非音近字或形近字字符替换操作、交换字符操作，那么子状态就由这五种操作方式转移得到现在状态。

[0056] 参见图2，是本发明一个实施例提供的一种加权编辑距离计算方法流程图，包括：

[0057] S201：定义状态转移方程，用于表示搜索词与热词之间的加权编辑距离，其中，在状态转移方程中定义两个状态量，用于分别表示搜索词和热词之间对应位置的字符。

[0058] S202：根据为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作所设置的不同数值的权重，求解状态转移方程在相应操作的解，得到加权编辑距离。

[0059] 本发明实施例中，热词与错误搜索词之间的相似度采用一种加权编辑算法实现。首先，定义状态转移方程edit(i,j)，表示第一个字符串长度为i的子串到第二个字符串长

度为j的子串的加权编辑距离，str1(i)表示第一个字符串的第i+1个字符，str2(j)表示第二个字符串的第j+1个字符。下面，提供一种基于动态规划的加权编辑距离计算方法逻辑实现实例：

[0060] A.if i==0 and j==0,edit(i,j)=0；

[0061] B.if i==0 and j>0,edit(i,j)=j；

[0062] C.if i>0 and j==0,edit(i,j)=i；

[0063] D.ifi≥1 and j≥1,edit(i,j)=min{edit(i-1,j)+1,edit(I,j-1)+1,edit(i-1,j-1)+f(i,j)}，其中：

[0064] D1.if str1(i-1)==str2(j-1),f(i,j)=0

[0065] D2.if str1(i-1),str2(j-1)互为形近字或音近字,f(i,j)=w1

[0066] D3.if i≥2 and j≥2 and str1(i-2)==str2(j-1)and str1(i-1)==str2(j-2),f(i,j)=1-w2

[0067] D4.其他情况下,f(i,j)=1

[0068] 其中，步骤A,B,C初始化函数edit(i,j)，步骤D1表示字符str1(i-1)和字符str2(j-1)相同，步骤D2表示字符str1(i-1)和字符str2(j-1)互为形近字或者音近字的替换操作，步骤D3表示交换操作，步骤D4表示插入、删除以及非形近字或音近字的替换操作；f(i,j)为操作代价值，f(i,j)根据为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作所设置的不同数值的权重，得到各权重对应的代价值。由此可以计算错误搜索词和热词之间的编辑距离。

[0069] 在步骤D2中，可以通过预先设置的音近字映射表或形近字映射表，来判断搜索词与热词之间是否互为音近字或形近字。例如，对于汉字而言，音近字映射表可以首先提取汉字的拼音，然后找到该拼音所包含的所有汉字，从而建立音近字映射表。同理，对于非汉语的其他语种，也可以采取类似的方式建立映射表。

[0070] S103：根据加权编辑距离和热词热度，选取预定数目的热词进行纠错提示。

[0071] 具体的，选择推荐词可由以下步骤完成：1、将热词搜索次数进行归一化处理；2、根据加权编辑距离与热词搜索次数归一化处理结果，计算推荐综合得分；3、选择推荐综合得分最高且所述加权编辑距离小于预定值的预定数目的热词，作为纠错的推荐词。

[0072] 在选择推荐词时，需要综合考虑编辑距离和热词热度的影响。假设加权编辑距离为edit_score，热词搜索次数为impression_count，采用对数公式将热词搜索次数归一化到0-1之间，例如，归一化公式为：

[0073] hot_index=min(log(impression_count+1)/20,1)

[0074] 那么，推荐综合得分为：

[0075] final_score=hot_index*edit_score

[0076] 最后，选择推荐综合得分最高且加权编辑距离小于预定值的k个热词作为纠错提示的推荐词，进行纠错提示。

[0077] 需要说明的是，对于方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本发明实施例并不受所描述的动作顺序的限制，因为依据本发明实施例，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作并不一定是本发明实施

例所必须的。

[0078] 参照图3，是本发明实施例提供的一种搜索词纠错装置结构示意图。该装置包括：

[0079] 错误搜索词识别单元301，用于识别出错误的搜索词；

[0080] 对搜索词进行纠错，是针对错误的搜索词进行纠错，因此首先需要识别出错误的搜索词。在具体实现中，可以基于搜索日志，识别出错误的搜索词：通过搜索词点击率、搜索结果完全匹配的结果个数、以及基于语言模型的搜索词概率，可有效识别出错误的搜索词。

[0081] 加权编辑距离计算单元302，用于利用加权编辑距离算法，计算所述搜索词与预先获取的热词之间的加权编辑距离，其中，在所述加权编辑距离计算过程中，针对从搜索词转换到热词的操作，分别为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作，设置不同数值的权重；

[0082] 热词，是指网络或当下常用或流行的词，在实际操作中，可通过点击率确定众多热词组成热词表。在热词表的产生过程中，需要过滤掉点击率低、搜索结果少的词语，以保证热词的准确性。本发明实施例中，通过将错误的搜索词纠正为编辑距离相近的热词，从而为用户提供更好的体验。

[0083] 本领域技术人员理解，编辑距离(Edit Distance)，又称Levenshtein距离，是指两个字串之间，由一个转成另一个所需的最少编辑操作次数。传统的编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。一般来说，编辑距离越小，两个字符串的相似度越大。

[0084] 传统的编辑距离操作包括：将一个字符替换成另一个字符，插入一个字符，删除一个字符，每一种操作对应的距离都是1。这些操作中并不包括字符串内临近字符之间的交换操作，也没有考虑音近字或形近字替换操作的特殊性。交换操作可以通过两次替换操作实现，比如CD->DD->DC，按照传统编辑距离算法，对应的距离是2。考虑到实际搜索过程中，用户将两个字符输入错位的概率非常高，交换操作的距离为2是不合理的。另外，因此，用户出现音近字或形近字导致的搜索词输入错误的概率也较高，如果对此特殊的替换操作没有特别处理，也设置距离为1，显然也是不合理的。

[0085] 因此，本发明实施例中提出了一种加权编辑距离方法，包括以下5种操作，并分别设置不同的权重。

[0086] 1)插入字符操作，权重为1；

[0087] 2)删除字符操作，权重为1；

[0088] 3)非形近字或者音近字的替换操作，权重为1；

[0089] 4)形近字或音近字的替换操作，权重为w1；

[0090] 5)交换字符操作，权重为w2。

[0091] 纠错提示单元303，用于根据所述加权编辑距离和热词热度，选取预定数目的热词作为纠错推荐词。

[0092] 优选的，所述加权编辑距离计算单元302包括：

[0093] 状态转移方程定义子单元3021，用于定义状态转移方程，用于表示所述搜索词与热词之间的加权编辑距离，其中，在状态转移方程中定义两个状态量，用于分别表示搜索词和热词之间对应位置的字符；

[0094] 方程求解子单元3022，用于根据为插入字符操作、删除字符操作、形近字或音近字

的替换操作、非形近字或音近字的替换操作、交换字符操作所设置的不同数值的权重，求解所述状态转移方程在相应操作的解，作为所述加权编辑距离。

[0095] 优选的，所述状态转移方程为：

[0096]

$$edit(i,j)=\min\{edit(i-1,j)+1,edit(i,j-1)+1,edit(i-1,j-1)+f(i,j),\};$$

[0097] 其中，i、j为所述两个状态量，f(i,j)为操作代价值，f(i,j)根据为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作所设置的不同数值的权重，得到各权重对应的代价值。

[0098] 优选的，该装置还包括：

[0099] 形近字或音近字确定单元304，用于查找预先设置的形近字映射表或音近字映射表，确定所述搜索词与所述热词是否互为形近字或音近字。

[0100] 优选的，该装置还包括：

[0101] 操作权重设置单元305，用于设置各操作权重满足如下关系：

[0102] 形近字或音近字的替换操作权重<交换字符操作权重<插入字符操作权重＝删除字符操作权重＝非形近字或音近字的替换操作权重。为了得到更好的效果，w1和w2的取值需特别注意，一般而言，需要满足w1<w2<1。经过实验得到，优选的，设置各操作权重满足如下关系：形近字或音近字的替换操作权重<交换字符操作权重<插入字符操作权重＝删除字符操作权重＝非形近字或音近字的替换操作权重。由于上述考虑了各种可能的操作，并且分别赋予不同的权重，因此采取这种加权编辑距离应用于纠错，计算字符串之间的相似度，精度更高。

[0103] 优选的，所述错误搜索词识别单元301包括：

[0104] 日志查找及计算子单元3011，用于基于搜索日志，解析或计算出待识别搜索词的搜索点击率、词特征、出现概率、全匹配结果数和全匹配占比；

[0105] 具体的，首先，计算待识别搜索词的搜索点击率，例如，从搜索日志中获取用户针对待识别搜索词的搜索次数和点击搜索结果次数；将点击搜索结果次数除以搜索次数，得到搜索点击率；其次，对待识别搜索词进行分词处理，得到多个词特征；继而，利用统计语言模型和各个词特征，计算待识别搜索词出现的概率；然后，计算待识别搜索词的全匹配结果数和相关结果数，其中，全匹配结果数为针对待识别搜索词的所有搜索结果中包含待识别搜索词的全部内容的结果的个数，相关结果数为针对待识别搜索词的所有搜索结果中包含待识别搜索词的部分内容的结果的个数；最后，计算全匹配结果数和相关结果数的比值，得到全匹配占比。

[0106] 识别结果确定子单元3012，用于根据待识别搜索词的所述搜索点击率、所述词特征、所述出现概率、所述全匹配结果数和所述全匹配占比，确定所述待识别搜索词为错误搜索词或正常搜索词。

[0107] 通过融合待识别搜索词的多维度特征(即搜索点击率、各个所述词特征、待识别搜索词出现的概率、全匹配结果数和全匹配占比)，并基于多维度特征对待识别搜索词进行识别，降低了对待识别搜索词进行识别的难度，从而提高了对待识别搜索词的识别能力，有利于识别出待识别搜索词是否为错误搜索词。

[0108] 优选的，所述纠错提示单元303包括：

[0109] 归一化处理子单元3031，用于将热词搜索次数进行归一化处理；

[0110] 推荐综合得分计算子单元3032，用于根据所述加权编辑距离与热词搜索次数归一化处理结果，计算推荐综合得分；

[0111] 推荐词确定子单元3033，用于选择推荐综合得分最高且所述加权编辑距离小于预定值的预定数目的热词，作为纠错的推荐词，进行纠错提示。

[0112] 对于装置实施例而言，由于其与方法实施例基本相似，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0113] 本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。

[0114] 本领域内的技术人员应明白，本发明实施例的实施例可提供为方法、装置、或计算机程序产品。因此，本发明实施例可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明实施例可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0115] 本发明实施例是参照根据本发明实施例的方法、终端设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理终端设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理终端设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0116] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理终端设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0117] 这些计算机程序指令也可装载到计算机或其他可编程数据处理终端设备上，使得在计算机或其他可编程终端设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程终端设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0118] 尽管已描述了本发明实施例的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例做出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本发明实施例范围的所有变更和修改。

[0119] 最后，还需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者终端设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者终端设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要

素，并不排除在包括所述要素的过程、方法、物品或者终端设备中还存在另外的相同要素。

[0120] 以上对本发明所提供的一种关系型数据库的调度方法及系统，进行了详细介绍，本文中应用了具体个例对本发明的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本发明的方法及其核心思想；同时，对于本领域的一般技术人员，依据本发明的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本发明的限制。

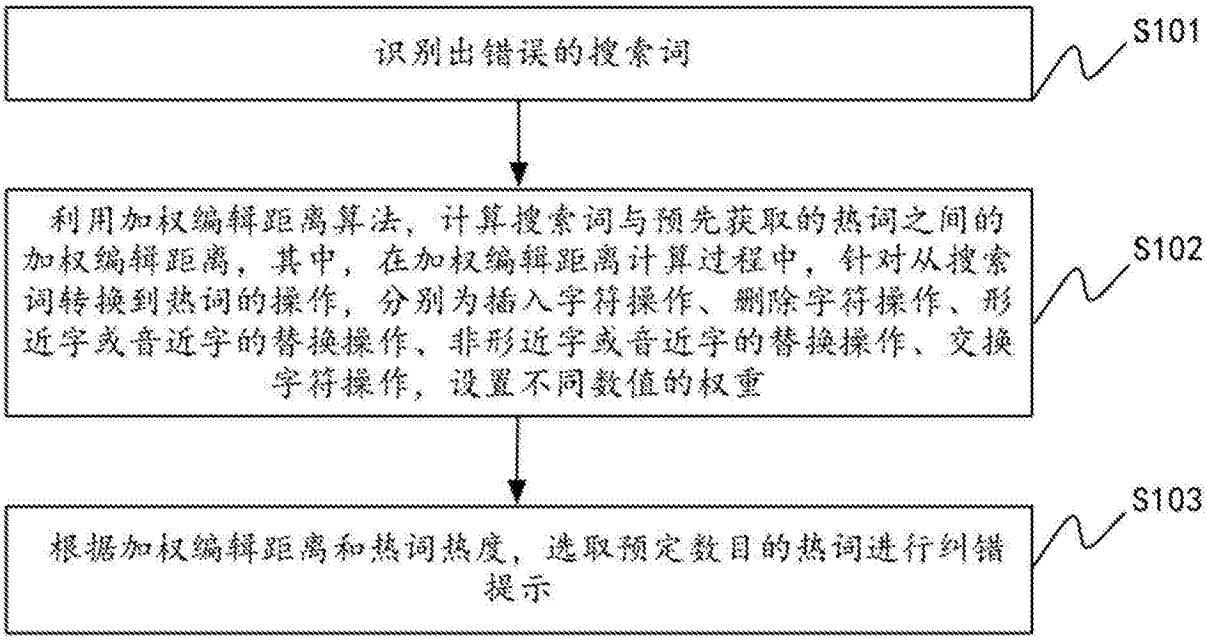

识别出错误的搜索词
S101
利用加权编辑距离算法，计算搜索词与预先获取的热词之间的加权编辑距离，其中，在加权编辑距离计算过程中，针对从搜索词转换到热词的操作，分别为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作，设置不同数值的权重
S102
根据加权编辑距离和热词热度，选取预定数目的热词进行纠错提示
S103

图1

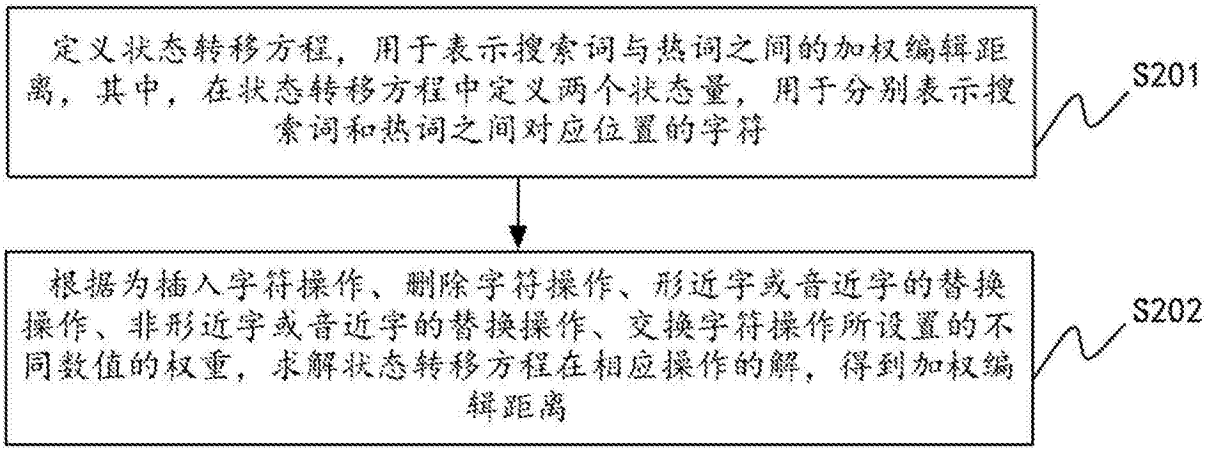

定义状态转移方程，用于表示搜索词与热词之间的加权编辑距离，其中，在状态转移方程中定义两个状态量，用于分别表示搜索词和热词之间对应位置的字符
S201
根据为插入字符操作、删除字符操作、形近字或音近字的替换操作、非形近字或音近字的替换操作、交换字符操作所设置的不同数值的权重，求解状态转移方程在相应操作的解，得到加权编辑距离
S202

图2

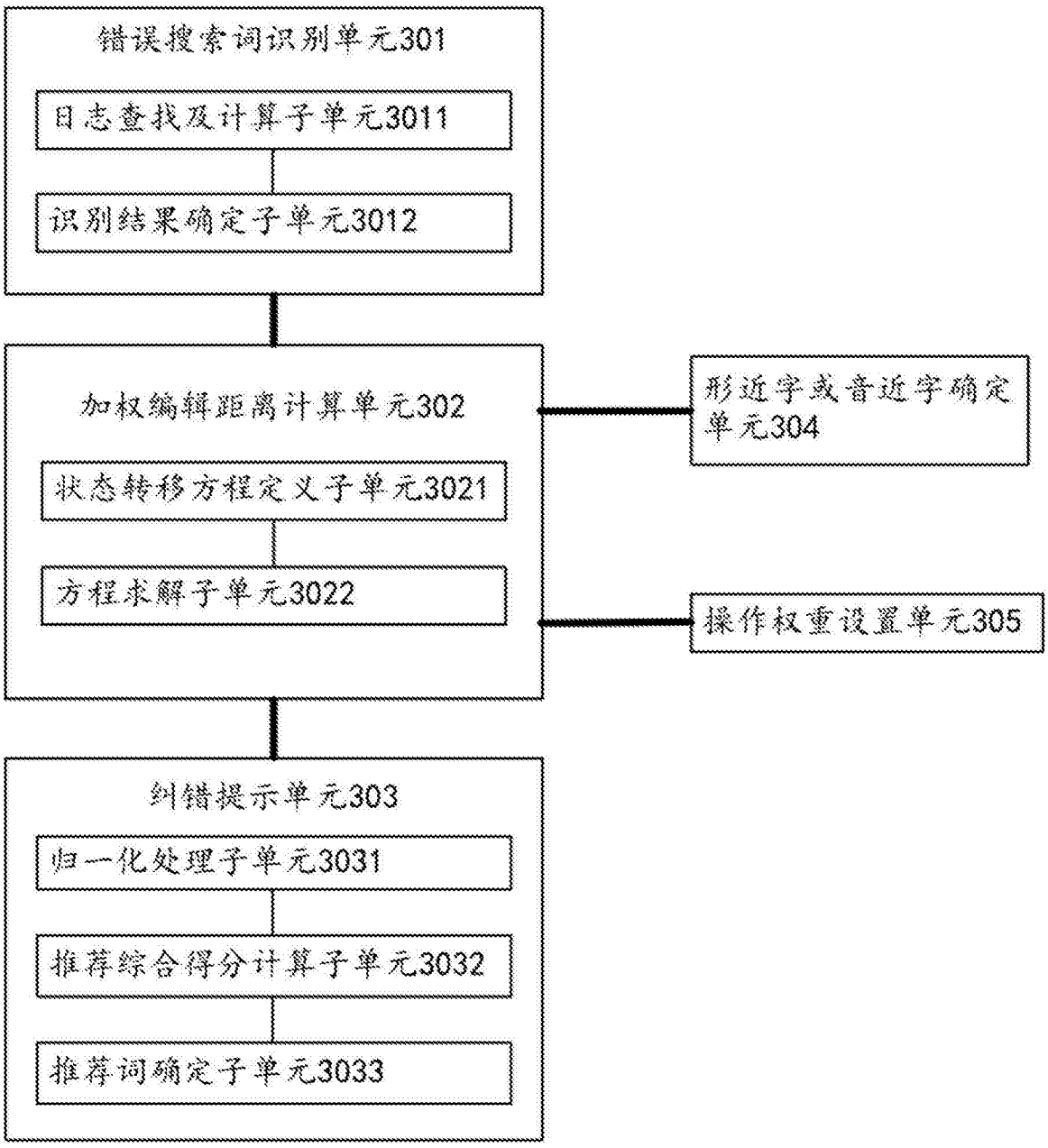

错误搜索词识别单元301
日志查找及计算子单元3011
识别结果确定子单元3012
加权编辑距离计算单元302
状态转移方程定义子单元3021
方程求解子单元3022
形近字或音近字确定单元304
操作权重设置单元305
纠错提示单元303
归一化处理子单元3031
推荐综合得分计算子单元3032
推荐词确定子单元3033

图3