

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



(10) 国际公布号

WO 2020/098098 A1

(43) 国际公布日

2020 年 5 月 22 日 (22.05.2020)

(51) 国际专利分类号:

G06F 17/24 (2006.01)

(21) 国际申请号:

PCT/CN2018/124398

(22) 国际申请日:

2018 年 12 月 27 日 (27.12.2018)

(25) 申请语言:

中文

(26) 公布语言:

中文

(30) 优先权:

201811348583.1 2018 年 11 月 13 日 (13.11.2018) CN

(71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];

中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 吴建财(WU, Jiancai); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 邹芳(ZOU, Fang); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 邢艳(XING, Yan); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市明日今典知识产权代理事务所(普通合伙)(SHENZHEN MINGRIJINDIAN INTELLECTUAL PROPERTY AGENCY FIRM (GENERAL)); 中国广东省深圳市南山区南头街道智恒新兴产业园 E 区 01B 栋 405 室, Guangdong 518000 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG,

(54) Title: SEMANTIC ANALYSIS-BASED TEXT ACCURACY CALCULATION METHOD, DEVICE AND COMPUTER DEVICE

(54) 发明名称: 基于语义解析的文本准确率计算方法、装置、计算机设备

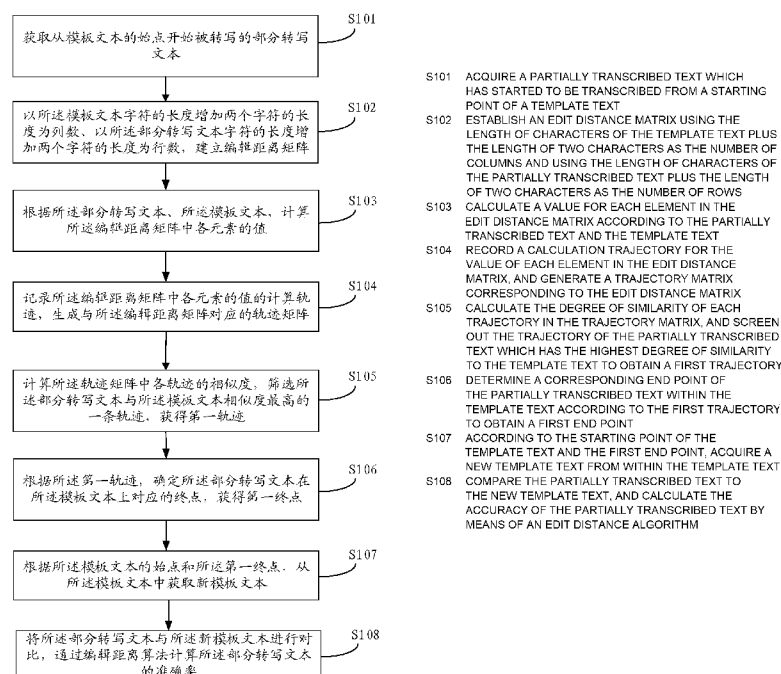


图 1

(57) Abstract: The present application relates to a semantic analysis-based text accuracy calculation method, device and computer device. When starting to transcribe text at a starting point of a template text, generating a trajectory matrix according to a calculation trajectory for the value of each element in an edit distance matrix; calculating the degree of similarity of each trajectory in the trajectory matrix; screening out the trajectory having the highest degree of similarity to obtain a first trajectory; obtain a new template text, and compare the partially transcribed text to the new template text so as to calculate the accuracy of the partially transcribed text.

BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

- (84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

- 包括国际检索报告(条约第21条(3))。

(57) 摘要: 本申请涉及一种基于语义解析的文本准确率计算方法、装置、计算机设备。在模板文本的始点开始被转写时, 根据编辑距离矩阵中各元素的值的计算轨迹生成轨迹矩阵, 计算轨迹矩阵中各轨迹的相似度, 筛选相似度最高的一条轨迹获得第一轨迹, 获得新模板文本, 再将部分转写文本与新模板文本进行对比, 计算部分转写文本的准确率。

基于语义解析的文本准确率计算方法、装置、计算机设备

本申请要求于 2018 年 11 月 13 日提交中国专利局、申请号为 2018113485831，申请名称为“基于语义解析的文本准确率计算方法、装置、计算机设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5

技术领域

本申请涉及语义解析技术领域，特别涉及一种基于语义解析的文本准确率计算方法、装置、计算机设备。

10 背景技术

在统计 ASR（语音识别）引擎转写准确率的时候，常用的算法是编辑距离算法。该算法通过统计转写文本转成模板文本所需的最少编辑操作（编辑操作包括：将一个字符替换成另外一个字符，插入一个字符，删除一个字符）次数来计算转写文本与模板文本的相似度（转写准确率）。但在关注 ASR 引擎的实时转写准确率的场景下，该算法的计算结果并不能令人满意。发明人意识到，由于该算法总是拿已经转写出来的文本与模板文本的全部文本进行对比，因此，当只有部分文本被转写出来的时候，该算法并不能准确地计算出这部分转写出来的文本的转写准确率。因此，编辑距离在关注 ASR 引擎实时转写准确率的场景下并不适用。

技术问题

20 针对现有技术不足，本申请提出一种基于语义解析的文本准确率计算方法、装置、计算机设备，旨在解决现有的文本的转写准确率算法，将已经转写出来的文本与模板文本的全部文本进行对比，在部分文本被转写出来的时候，不能准确计算文本的转写准确率的问题。

技术解决方案

25 本申请提出的技术方案是：

一种基于语义解析的文本准确率计算方法，所述方法包括：

获取从模板文本的始点开始被转写的部分转写文本；

以所述模板文本字符的长度增加两个字符的长度为列数、以所述部分转写文本字符的长度增加两个字符的长度为行数，建立编辑距离矩阵；

根据所述部分转写文本、所述模板文本，计算所述编辑距离矩阵中各元素的值；

记录所述编辑距离矩阵中各元素的值的计算轨迹，生成与所述编辑距离矩阵对应的轨迹矩阵；

计算所述轨迹矩阵中各轨迹的相似度，筛选所述部分转写文本与所述模板文本相似度最高的一条轨迹，获得第一轨迹；

5 根据所述第一轨迹，确定所述部分转写文本在所述模板文本上对应的终点，获得第一终点；

根据所述模板文本的始点和所述第一终点，从所述模板文本中获取新模板文本；

将所述部分转写文本与所述新模板文本进行对比，通过编辑距离算法计算所述部分转写文本的准确率。

本申请还提供一种基于语义解析的文本准确率计算装置，所述装置包括：

10 第一获取模块，用于获取从模板文本的始点开始被转写的部分转写文本；

建立模块，用于以所述模板文本字符的长度增加两个字符的长度为列数、以所述部分转写文本字符的长度增加两个字符的长度为行数，建立编辑距离矩阵；

第一计算模块，用于根据所述部分转写文本、所述模板文本，计算所述编辑距离矩阵中各元素的值；

生成模块，用于记录所述编辑距离矩阵中各元素的值的计算轨迹，生成与所述编辑距离矩阵对应的

15 轨迹矩阵；

筛选模块，用于计算所述轨迹矩阵中各轨迹的相似度，筛选所述部分转写文本与所述模板文本相似度最高的一条轨迹，获得第一轨迹；

获得模块，用于根据所述第一轨迹，确定所述部分转写文本在所述模板文本上对应的终点，获得第一终点；

20 第二获取模块，用于根据所述模板文本的始点和所述第一终点，从所述模板文本中获取新模板文本；

第二计算模块，用于将所述部分转写文本与所述新模板文本进行对比，通过编辑距离算法计算所述部分转写文本的准确率。

本申请还提供一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述处理器执行所述计算机程序时实现上述任一项所述的方法的步骤。

25 本申请还提供一种计算机非易失性可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现上述任一项所述的方法的步骤。

有益效果

根据上述的技术方案，本申请有益效果：在模板文本的始点开始被转写时，建立编辑距离矩阵，计算编

30 辑距离矩阵中各元素的值，根据编辑距离矩阵中各元素的值的计算轨迹生成轨迹矩阵，计算轨迹矩阵中

各轨迹的相似度，筛选相似度最高的一条轨迹获得第一轨迹，根据第一轨迹获得部分转写文本在模板文本上对应的终点，从而获得新模板文本，再将部分转写文本与新模板文本进行对比，计算部分转写文本的准确率，旨在解决现有的文本的转写准确率算法，将已经转写出来的文本与模板文本的全部文本进行对比，在部分文本被转写出来的时候，不能准确计算文本的转写准确率的问题。

5 附图说明

图1是应用本申请实施例提供的基于语义解析的文本准确率计算方法的流程图；

图2是应用本申请实施例提供的基于语义解析的文本准确率计算装置的功能模块图；

图3是应用本申请实施例提供的计算机设备的结构示意图。

本发明的最佳实施方式

10 如图1所示，本申请实施例提出一种基于语义解析的文本准确率计算方法，所述方法包括以下步骤：

步骤 S101、获取从模板文本的始点开始被转写的部分转写文本。

从模板文本的始点开始被转写，且模板文本未全部被转写，也就是从模板文本的第一个字符开始被转写，但是转写的结束点不是在模板文本的最后一个字符，而是在除了模板文本的最后一个字符之外，在模板文本中任意一个字符。由于不是对模板文本全部的字符的转写，为此，从模板文本的始点开始被
15 转写所得到文本称为部分转写文本。

模板文本是一个正确的文本，用于与部分转写文本进行对比的文本。

上述的转写是指通过 ASR（语音识别）引擎将语音转写为文本。

步骤 S102、以所述模板文本字符的长度增加两个字符的长度为列数、以所述部分转写文本字符的长度增加两个字符的长度为行数，建立编辑距离矩阵。

20 在本实施例中，模板文本为剔除标点符号的文字文本。部分转写文本为剔除标点符号的文字文本。

获取模板文本字符的长度，根据模板文本字符长度再增加两个字符的长度，作为列数。获取部分转写文本字符的长度，根据部分转写文本字符的长度再增加两个字符的长度，作为行数，然后以模板文本字符的长度增加两个字符的长度为列数、以部分转写文本字符的长度增加两个字符的长度为行数，建立编辑距离矩阵。模板文本字符长度再增加两个字符的长度作为列数、部分转写文本字符的长度再增加两
25 个字符的长度作为行数的目的是为了在第一行、第一列上分别输入模板文本、部分转写文本，以及在第二行、第二列输入初始化的值。

具体地，在步骤 S102 之后，且在步骤 S103 之前，包括：

从所述编辑距离矩阵的第一行的第三个元素开始输入所述模板文本的字符；

从所述编辑距离矩阵的第一列的第三个元素开始输入所述部分转写文本的字符；

30 定义所述编辑距离矩阵的第二行中的第二个元素的值为 0；

以所述编辑距离矩阵的第二行中的第二个元素的值为 0 依次递增数值 1，初始化所述编辑距离矩阵

的第二行的各元素的值;

以所述编辑距离矩阵的第二列中的第二个元素的值为 0 依次递增数值 1, 初始化所述编辑距离矩阵的第二列的各元素的值。

在编辑距离矩阵的第一行中输入模板文本的字符, 具体地, 从编辑距离矩阵的第一行的第三个元素
5 开始输入模板文本的字符。对应地, 在编辑距离矩阵的第一列中输入部分转写文本的字符, 具体地, 从编辑距离矩阵的第一列的第三个元素开始输入部分转写文本的字符。编辑距离矩阵的第一行、第一列的第三个元素分别开始输入模板文本的字符、部分转写文本的字符, 使模板文本的各字符与部分转写文本各字符都在编辑距离矩阵存在对应关系, 另外, 也是为了对第二行、第二列的初始化的数值提供对应的位置关系。首先, 定义编辑距离矩阵的第二行中的第二个元素的值为 0, 然后, 以编辑距离矩阵的第二
10 行中的第二个元素的值为 0 依次递增数值 1, 初始化编辑距离矩阵的第二行的各元素的值, 例如, 编辑距离矩阵的第二行中的第二、三、四、五个元素的值分别为 0、1、2、3。定义编辑距离矩阵的第二行中的第二个元素的值为 0, 实质上, 也定义编辑距离矩阵的第二列中的第二个元素的值为 0, 因为编辑距离矩阵的第二行中的第二个元素与编辑距离矩阵的第二列中的第二个元素是在同一个位置, 即使同一个元素, 以编辑距离矩阵的第二列中的第二个元素的值为 0 依次递增数值 1, 初始化编辑距离矩阵的第二
15 列的各元素的值, 例如, 编辑距离矩阵的第二列中的第二、三、四、五个元素的值分别为 0、1、2、3。在初始化编辑距离矩阵的第二列、第二行的数值之后, 可能进行计算编辑距离矩阵中各元素的值。

步骤 S103、根据所述部分转写文本、所述模板文本, 计算所述编辑距离矩阵中各元素的值。

根据部分转写文本、模板文本, 也就是, 在编辑距离矩阵中, 部分转写文本的字符对应模板文本的字符是否相等, 决定编辑距离矩阵中各元素的值的计算方式, 进而计算编辑距离矩阵中各元素的值。

20 在本实施例中, 编辑距离矩阵中未被初始化的各元素的值由其左方、左上角、上方中的某一个元素的值来确定。在步骤 S103 中, 包括:

识别所述编辑距离矩阵的第三列中的第三个元素所处的列数、行数;

识别所述编辑距离矩阵的第三列中的第三个元素所处的列数、行数分别对应所述模板文本的字符、所述部分转写文本的字符;

25 判断所述编辑距离矩阵的第三列中的第三个元素所处的列数对应所述模板文本的字符与所述编辑距离矩阵的第三列中的第三个元素处于行数对应所述部分转写文本的字符是否相等;

若所述编辑距离矩阵的第三列中的第三个元素所处的列数对应所述模板文本的字符与所述编辑距离矩阵的第三列中的第三个元素处于行数对应所述部分转写文本的字符相等, 则所述编辑距离矩阵的第三列中的第三个元素的值为其左上角的元素的值;

30 若所述编辑距离矩阵的第三列中的第三个元素所处的列数对应所述模板文本的字符与所述编辑距离矩阵的第三列中的第三个元素处于行数对应所述部分转写文本的字符不相等, 则所述编辑距离矩

阵的第三列中的第三个元素的值为其左方、左上角、上方的元素中最小值加1得到;

依次计算所述编辑距离矩阵的第三列中的第四个元素的值,直至完成计算所述编辑距离矩阵中各元素的值。

由于编辑距离矩阵中未被初始化的各元素的值由其左方、左上角、上方中的某一个元素的值来确定, 5 在开始计算时,符合左方、左上角、上方的元素都存在数值的只有编辑距离矩阵的第三列中的第三个元素,或者说编辑距离矩阵的第三列中的第三个元素,在本实施例中,计算编辑距离矩阵的第三列中的第三个元素,识别编辑距离矩阵的第三列中的第三个元素所处的列数、行数,在获得编辑距离矩阵的第三列中的第三个元素所处的列数、行数之后,识别编辑距离矩阵的第三列中的第三个元素所处的列数对应模板文本的字符,编辑距离矩阵的第三列中的第三个元素所处的行数对应部分转写文本的字符。 10 在获得对应的字符之后,判断编辑距离矩阵的第三列中的第三个元素所处的列数对应模板文本的字符与编辑距离矩阵的第三列中的第三个元素处于行数对应部分转写文本的字符是否相等,根据编辑距离矩阵的第三列中的第三个元素的模板文本的字符与对应的部分转写文本的字符是否相等,用于确定编辑距离矩阵的第三列中的第三个元素的值,若编辑距离矩阵的第三列中的第三个元素所处的列数对应模板文本的字符与编辑距离矩阵的第三列中的第三个元素处于行数对应部分转写文本的字符相等, 15 则编辑距离矩阵的第三列中的第三个元素的值为其左上角的元素的值。若编辑距离矩阵的第三列中的第三个元素所处的列数对应模板文本的字符与编辑距离矩阵的第三列中的第三个元素处于行数对应部分转写文本的字符不相等,则编辑距离矩阵的第三列中的第三个元素的值为其左方、左上角、上方的元素中最小值加1得到。在计算完编辑距离矩阵的第三列中的第三个元素的值之后,依次计算编辑距离矩阵的第三列中的第四个元素的值,直至完成计算编辑距离矩阵中各元素的值,也就是,接着计算编辑距离矩阵的第三列中的第四个元素的值,在计算完编辑距离矩阵的第三列各元素的值,再计算编辑距离 20 矩阵的第四列中各元素的值,直至计算完成编辑距离矩阵的最后一列中各元素的值,才完成计算编辑距离矩阵中各元素的值。

步骤S104、记录所述编辑距离矩阵中各元素的值的计算轨迹,生成与所述编辑距离矩阵对应的轨迹矩阵。

25 在计算编辑距离矩阵中各元素的值的过程中,记录编辑距离矩阵中各元素的值的计算轨迹,也就是,编辑距离矩阵中各元素的值是由哪一个元素的值决定的。在完成计算编辑距离矩阵中各元素的值之后,记录编辑距离矩阵中各元素的值的计算轨迹也完成,从而生成与编辑距离矩阵对应的轨迹矩阵。

在本实施例中,在步骤S104中,包括:

记录所述编辑距离矩阵中各元素的值的计算轨迹;

30 根据所述编辑距离矩阵中各元素的值的计算轨迹,标记所述编辑距离矩阵中各元素的值产生由来;在完成标记后,生成与所述编辑距离矩阵对应的轨迹矩阵。

记录编辑距离矩阵中各元素的值的计算轨迹，根据编辑距离矩阵中各元素的值的计算轨迹，标记编辑距离矩阵中各元素的值产生由来，在本实施例中，用 lt 表示该元素通过左上方的元素计算而来，用 l 表示该元素通过左方的元素计算而来，用 t 表示该元素通过上方的元素计算而来，例如，若编辑距离矩阵的第三列中的第三个元素是由编辑距离矩阵的第二列中的第二个元素确定的，则在编辑距离矩阵的第三列中的第三个元素输入 lt ，若编辑距离矩阵的第三列中的第三个元素是由编辑距离矩阵的第二列中的第三个元素确定的，则在编辑距离矩阵的第三列中的第三个元素输入 l ，若编辑距离矩阵的第三列中的第三个元素是由编辑距离矩阵的第三列中的第二个元素确定的，则在编辑距离矩阵的第三列中的第三个元素输入 t ，从而标记编辑距离矩阵的第三列中的第三个元素的产生由来。在完成标记后，生成与编辑距离矩阵对应的轨迹矩阵。

在本实施例中，在所述根据所述编辑距离矩阵中各元素的值的计算轨迹，标记所述编辑距离矩阵中各元素的值产生由来的步骤中，包括：

在每记录所述编辑距离矩阵中一个元素的值的计算轨迹时，标记所述编辑距离矩阵中该元素的值产生由来；

直至标记所述编辑距离矩阵中各元素的值产生由来。

每记录编辑距离矩阵中一个元素的值的计算轨迹，就立刻标记编辑距离矩阵中该元素的值产生由来，也就是，一边记录编辑距离矩阵中各元素的值的计算轨迹，一边标记编辑距离矩阵中各元素的值产生由来。

在一些实施例中，在所述根据所述编辑距离矩阵中各元素的值的计算轨迹，标记所述编辑距离矩阵中各元素的值产生由来的步骤中，包括：

在完成记录所述编辑距离矩阵中各元素的值的计算轨迹之后，根据所述编辑距离矩阵中各元素的值的计算轨迹，标记所述编辑距离矩阵中各元素的值产生由来。

在完成编辑距离矩阵中各元素的值的计算轨迹之后，才会触发开始执行标记编辑距离矩阵中各元素的值产生由来，直至完成标记编辑距离矩阵中各元素的值产生由来。也就是，在未完成编辑距离矩阵中各元素的值的计算轨迹，不会进行标记编辑距离矩阵中各元素的值产生由来。

步骤 S105、计算所述轨迹矩阵中各轨迹的相似度，筛选所述部分转写文本与所述模板文本相似度最高的一条轨迹，获得第一轨迹。

在生成轨迹矩阵之后，计算轨迹矩阵中各轨迹的相似度，在完成计算各轨迹的相似度之后，筛选部分转写文本与模板文本相似度最高的一条轨迹，获得第一轨迹，该第一轨迹认为是部分转写文本在模板文本上对应轨迹。

在本实施例中，在所述计算所述轨迹矩阵中各轨迹的相似度的步骤中，包括：

识别所述轨迹矩阵中各轨迹中所述部分转写文本的字符与对应的所述模板文本的字符相等的个数，

获得相等字符个数;

比较所述轨迹矩阵中各轨迹中所述部分转写文本的字符的长度与对应的所述模板文本的字符的长度,选取长度长的作为字符总数;

5 计算所述轨迹矩阵中各轨迹的相等字符个数与对应的字符总数的比值,获得所述轨迹矩阵中各轨迹的相似度。

在生成轨迹矩阵之后,识别轨迹矩阵中各轨迹中部分转写文本的字符与对应的模板文本的字符相等的个数,获得相等字符个数,在获得轨迹矩阵中各轨迹中相等字符个数之后,比较轨迹矩阵中各轨迹中部分转写文本的字符的长度与对应的模板文本的字符的长度,选取长度长的作为字符总数,若轨迹矩阵中一轨迹中部分转写文本的字符的长度大于对应的模板文本的字符的长度,则在轨迹矩阵中一轨迹中选取部分转写文本的字符的长度为字符总数。若轨迹矩阵中一轨迹中部分转写文本的字符的长度小于对应的模板文本的字符的长度,则在轨迹矩阵中一轨迹中选取模板文本的字符的长度为字符总数。在选取长度长的作为字符总数之后,计算轨迹矩阵中各轨迹的相等字符个数与对应的字符总数的比值,在完成计算比值之后,获得轨迹矩阵中各轨迹的相似度。

15 步骤 S106、根据所述第一轨迹,确定所述部分转写文本在所述模板文本上对应的终点,获得第一终点。

在获得第一轨迹之后,根据第一轨迹,由于第一轨迹在轨迹矩阵中存在终点,确定部分转写文本在模板文本上对应的终点,从而获得第一终点。

在本实施例中,在步骤 S106 中,包括:

标记所述第一轨迹中最后一个元素;

20 根据所述第一轨迹中最后一个元素,标记对应的所述模板文本的字符,获得第一终点。

在获得第一轨迹之后,标记第一轨迹中最后一个元素,根据第一轨迹中最后一个元素,获得第一轨迹中最后一个元素对应列上的模板文本的字符,标记对应的模板文本的字符,从而获得第一终点。

步骤 S107、根据所述模板文本的始点和所述第一终点,从所述模板文本中获取新模板文本。

25 在获得第一终点之后,根据模板文本的始点和第一终点,获得两点之间的文本,包括模板文本的始点、第一终点分别对应的字符,从而从模板文本中获取新模板文本。

在本实施例中,在步骤 S107 中,包括:

在所述模板文本中标记所述模板文本的第一个字符为始点;

截取所述模板文本的始点与所述第一终点之间的字符,其中所述模板文本的始点与所述第一终点之间的字符包括所述模板文本的始点对应的字符和所述第一终点对应的字符;

30 根据截取到的字符生成文本,获取所述新模板文本。

在得到第一终点之后,在模板文本中标记模板文本的第一个字符为始点,截取模板文本的始点与第

一终点之间的字符，其中模板文本的始点与第一终点之间的字符包括模板文本的始点对应的字符和第一终点对应的字符。在截取到的字符之后，根据截取到的字符，生成与模板文本同样格式的文本，获得新模板文本。

步骤 S108、将所述部分转写文本与所述新模板文本进行对比，通过编辑距离算法计算所述部分转写文本的准确率。

在获得新模板文本之后，将部分转写文本与新模板文本进行对比，并不是与模板文本进行对比，通过编辑距离算法计算部分转写文本的准确率，从而解决现有的文本的转写准确率算法，将已经转写出来的文本与模板文本的全部文本进行对比，在部分文本被转写出来的时候，不能准确计算文本的转写准确率的问题。

综上所述，在模板文本的始点开始被转写时，建立编辑距离矩阵，计算编辑距离矩阵中各元素的值，根据编辑距离矩阵中各元素的值的计算轨迹生成轨迹矩阵，计算轨迹矩阵中各轨迹的相似度，筛选相似度最高的一条轨迹获得第一轨迹，根据第一轨迹获得部分转写文本在模板文本上对应的终点，从而获得新模板文本，再将部分转写文本与新模板文本进行对比，计算部分转写文本的准确率，旨在解决现有的文本的转写准确率算法，将已经转写出来的文本与模板文本的全部文本进行对比，在部分文本被转写出来的时候，不能准确计算文本的转写准确率的问题。

如图 2 所示，本申请实施例提出一种基于语义解析的文本准确率计算装置 1，装置 1 包括第一获取模块 11、建立模块 12、第一计算模块 13、生成模块 14、筛选模块 15、获得模块 16、第二获取模块 17 和第二计算模块 18。

第一获取模块 11，用于获取从模板文本的始点开始被转写的部分转写文本。

从模板文本的始点开始被转写，且模板文本未全部被转写，也就是从模板文本的第一个字符开始被转写，但是转写的结束点不是在模板文本的最后一个字符，而是在除了模板文本的最后一个字符之外，在模板文本中任意一个字符。由于不是对模板文本全部的字符的转写，为此，从模板文本的始点开始被转写所得到文本称为部分转写文本。

模板文本是一个正确的文本，用于与部分转写文本进行对比的文本。

上述的转写是指通过 ASR（语音识别）引擎将语音转写为文本。

建立模块 12，用于以所述模板文本字符的长度增加两个字符的长度为列数、以所述部分转写文本字符的长度增加两个字符的长度为行数，建立编辑距离矩阵。

在本实施例中，模板文本为剔除标点符号的文字文本。部分转写文本为剔除标点符号的文字文本。

获取模板文本字符的长度，根据模板文本字符长度再增加两个字符的长度，作为列数。获取部分转写文本字符的长度，根据部分转写文本字符的长度再增加两个字符的长度，作为行数，然后以模板文本字符的长度增加两个字符的长度为列数、以部分转写文本字符的长度增加两个字符的长度为行数，建立

编辑距离矩阵。模板文本字符长度再增加两个字符的长度作为列数、部分转写文本字符的长度再增加两个字符的长度作为行数的目的是为了在第一行、第一列上分别输入模板文本、部分转写文本，以及在第二行、第二列输入初始化的值。

具体地，装置 1 包括：

- 5 第一输入模块，用于从所述编辑距离矩阵的第一行的第三个元素开始输入所述模板文本的字符；
 第二输入模块，用于从所述编辑距离矩阵的第一列的第三个元素开始输入所述部分转写文本的字符；

 定义模块，用于定义所述编辑距离矩阵的第二行中的第二个元素的值为 0；

- 第一初始化模块，用于以所述编辑距离矩阵的第二行中的第二个元素的值为 0 依次递增数值 1，初
10 始化所述编辑距离矩阵的第二行的各元素的值；

 第二初始化模块，用于以所述编辑距离矩阵的第二列中的第二个元素的值为 0 依次递增数值 1，初始化所述编辑距离矩阵的第二列的各元素的值。

- 在编辑距离矩阵的第一行中输入模板文本的字符，具体地，从编辑距离矩阵的第一行的第三个元素开始输入模板文本的字符。对应地，在编辑距离矩阵的第一列中输入部分转写文本的字符，具体地，从
15 编辑距离矩阵的第一列的第三个元素开始输入部分转写文本的字符。编辑距离矩阵的第一行、第一列的第三个元素分别开始输入模板文本的字符、部分转写文本的字符，使模板文本的各字符与部分转写文本各字符都在编辑距离矩阵存在对应关系，另外，也是为了对第二行、第二列的初始化的数值提供对应的位置关系。首先，定义编辑距离矩阵的第二行中的第二个元素的值为 0，然后，以编辑距离矩阵的第二行中的第二个元素的值为 0 依次递增数值 1，初始化编辑距离矩阵的第二行的各元素的值，例如，编辑
20 距离矩阵的第二行中的第二、三、四、五个元素的值分别为 0、1、2、3。定义编辑距离矩阵的第二行中的第二个元素的值为 0，实质上，也定义编辑距离矩阵的第二列中的第二个元素的值为 0，因为编辑距离矩阵的第二行中的第二个元素与编辑距离矩阵的第二列中的第二个元素是在同一个位置，即使同一个元素，以编辑距离矩阵的第二列中的第二个元素的值为 0 依次递增数值 1，初始化编辑距离矩阵的第二列的各元素的值，例如，编辑距离矩阵的第二列中的第二、三、四、五个元素的值分别为 0、1、2、3。
25 在初始化编辑距离矩阵的第二列、第二行的数值之后，可能进行计算编辑距离矩阵中各元素的值。

 第一计算模块 13，用于根据所述部分转写文本、所述模板文本，计算所述编辑距离矩阵中各元素的值。

 根据部分转写文本、模板文本，也就是，在编辑距离矩阵中，部分转写文本的字符对应模板文本的字符是否相等，决定编辑距离矩阵中各元素的值的计算方式，进而计算编辑距离矩阵中各元素的值。

- 30 在本实施例中，编辑距离矩阵中未被初始化的各元素的值由其左方、左上角、上方中的某一个元素的值来确定。第一计算模块 13 包括：

第一识别模块,用于识别所述编辑距离矩阵的第三列中的第三个元素所处的列数、行数;

第二识别模块,用于识别所述编辑距离矩阵的第三列中的第三个元素所处的列数、行数分别对应所述模板文本的字符、所述部分转写文本的字符;

第一判断模块,用于判断所述编辑距离矩阵的第三列中的第三个元素所处的列数对应所述模板文本的字符与所述编辑距离矩阵的第三列中的第三个元素处于行数对应所述部分转写文本的字符是否相等;若所述编辑距离矩阵的第三列中的第三个元素所处的列数对应所述模板文本的字符与所述编辑距离矩阵的第三列中的第三个元素处于行数对应所述部分转写文本的字符相等,则所述编辑距离矩阵的第三列中的第三个元素的值为其左上角的元素的值;若所述编辑距离矩阵的第三列中的第三个元素所处的列数对应所述模板文本的字符与所述编辑距离矩阵的第三列中的第三个元素处于行数对应所述部分转写文本的字符不相等,则所述编辑距离矩阵的第三列中的第三个元素的值为其左方、左上角、上方的元素中最小值加1得到;

第一子计算模块,用于依次计算所述编辑距离矩阵的第三列中的第四个元素的值,直至完成计算所述编辑距离矩阵中各元素的值。

由于编辑距离矩阵中未被初始化的各元素的值由其左方、左上角、上方中的某一个元素的值来确定,在开始计算时,符合左方、左上角、上方的元素都存在数值的只有编辑距离矩阵的第三列中的第三个元素,或者说编辑距离矩阵的第三列中的第三个元素,在本实施例中,计算编辑距离矩阵的第三列中的第三个元素,识别编辑距离矩阵的第三列中的第三个元素所处的列数、行数,在获得编辑距离矩阵的第三列中的第三个元素所处的列数、行数之后,识别编辑距离矩阵的第三列中的第三个元素所处的列数对应模板文本的字符,编辑距离矩阵的第三列中的第三个元素所处的行数对应部分转写文本的字符。在获得对应的字符之后,判断编辑距离矩阵的第三列中的第三个元素所处的列数对应模板文本的字符与编辑距离矩阵的第三列中的第三个元素处于行数对应部分转写文本的字符是否相等,根据编辑距离矩阵的第三列中的第三个元素的模板文本的字符与对应的部分转写文本的字符是否相等,用于确定编辑距离矩阵的第三列中的第三个元素的值,若编辑距离矩阵的第三列中的第三个元素所处的列数对应模板文本的字符与编辑距离矩阵的第三列中的第三个元素处于行数对应部分转写文本的字符相等,则编辑距离矩阵的第三列中的第三个元素的值为其左上角的元素的值。若编辑距离矩阵的第三列中的第三个元素所处的列数对应模板文本的字符与编辑距离矩阵的第三列中的第三个元素处于行数对应部分转写文本的字符不相等,则编辑距离矩阵的第三列中的第三个元素的值为其左方、左上角、上方的元素中最小值加1得到。在计算完编辑距离矩阵的第三列中的第三个元素的值之后,依次计算编辑距离矩阵的第三列中的第四个元素的值,直至完成计算编辑距离矩阵中各元素的值,也就是,接着计算编辑距离矩阵的第三列中的第四个元素的值,在计算完编辑距离矩阵的第三列各元素的值,再计算编辑距离矩阵的第四列中各元素的值,直至计算完成编辑距离矩阵的最后一列中各元素的值,才完成计算编辑距

离矩阵中各元素的值。

生成模块 14, 用于记录所述编辑距离矩阵中各元素的值的计算轨迹, 生成与所述编辑距离矩阵对应的轨迹矩阵。

在计算编辑距离矩阵中各元素的值的过程中, 记录编辑距离矩阵中各元素的值的计算轨迹, 也就是, 编辑距离矩阵中各元素的值是由哪一个元素的值决定的。在完成计算编辑距离矩阵中各元素的值之后, 记录编辑距离矩阵中各元素的值的计算轨迹也完成, 从而生成与编辑距离矩阵对应的轨迹矩阵。

在本实施例中, 生成模块 14 包括:

第一记录模块, 用于记录所述编辑距离矩阵中各元素的值的计算轨迹;

第一标记模块, 用于根据所述编辑距离矩阵中各元素的值的计算轨迹, 标记所述编辑距离矩阵中各元素的值产生由来;

第一生成模块, 用于在完成标记后, 生成与所述编辑距离矩阵对应的轨迹矩阵。

记录编辑距离矩阵中各元素的值的计算轨迹, 根据编辑距离矩阵中各元素的值的计算轨迹, 标记编辑距离矩阵中各元素的值产生由来, 在本实施例中, 用 $1t$ 表示该元素通过左上方的元素计算而来, 用 1 表示该元素通过左方的元素计算而来, 用 t 表示该元素通过上方的元素计算而来, 例如, 若编辑距离矩阵的第三列中的第三个元素是由编辑距离矩阵的第二列中的第二个元素确定的, 则在编辑距离矩阵的第三列中的第三个元素输入 $1t$, 若编辑距离矩阵的第三列中的第三个元素是由编辑距离矩阵的第二列中的第三个元素确定的, 则在编辑距离矩阵的第三列中的第三个元素输入 1 , 若编辑距离矩阵的第三列中的第三个元素是由编辑距离矩阵的第三列中的第二个元素确定的, 则在编辑距离矩阵的第三列中的第三个元素输入 t , 从而标记编辑距离矩阵的第三列中的第三个元素的产生由来。在完成标记后, 生成与编辑距离矩阵对应的轨迹矩阵。

在本实施例中, 第一标记模块包括:

第一子标记模块, 用于在每记录所述编辑距离矩阵中一个元素的值的计算轨迹时, 标记所述编辑距离矩阵中该元素的值产生由来;

第一子标记完成模块, 用于直至标记所述编辑距离矩阵中各元素的值产生由来。

每记录编辑距离矩阵中一个元素的值的计算轨迹, 就立刻标记编辑距离矩阵中该元素的值产生由来, 也就是, 一边记录编辑距离矩阵中各元素的值的计算轨迹, 一边标记编辑距离矩阵中各元素的值产生由来。

在一些实施例中, 第一标记模块包括:

第二子标记模块, 用于在完成记录所述编辑距离矩阵中各元素的值的计算轨迹之后, 根据所述编辑距离矩阵中各元素的值的计算轨迹, 标记所述编辑距离矩阵中各元素的值产生由来。

在完成编辑距离矩阵中各元素的值的计算轨迹之后, 才会触发开始执行标记编辑距离矩阵中各元素

的值产生由来，直至完成标记编辑距离矩阵中各元素的值产生由来。也就是，在未完成编辑距离矩阵中各元素的值的计算轨迹，不会进行标记编辑距离矩阵中各元素的值产生由来。

筛选模块 15，用于计算所述轨迹矩阵中各轨迹的相似度，筛选所述部分转写文本与所述模板文本相似度最高的一条轨迹，获得第一轨迹。

- 5 在生成轨迹矩阵之后，计算轨迹矩阵中各轨迹的相似度，在完成计算各轨迹的相似度之后，筛选部分转写文本与模板文本相似度最高的一条轨迹，获得第一轨迹，该第一轨迹认为是部分转写文本在模板文本上对应轨迹。

在本实施例中，筛选模块 15 包括：

- 10 第三识别模块，用于识别所述轨迹矩阵中各轨迹中所述部分转写文本的字符与对应的所述模板文本的字符相等的个数，获得相等字符个数；

第一比较模块，用于比较所述轨迹矩阵中各轨迹中所述部分转写文本的字符的长度与对应的所述模板文本的字符的长度，选取长度长的作为字符总数；

第三计算模块，用于计算所述轨迹矩阵中各轨迹的相等字符个数与对应的字符总数的比值，获得所述轨迹矩阵中各轨迹的相似度。

- 15 在生成轨迹矩阵之后，识别轨迹矩阵中各轨迹中部分转写文本的字符与对应的模板文本的字符相等的个数，获得相等字符个数，在获得轨迹矩阵中各轨迹中相等字符个数之后，比较轨迹矩阵中各轨迹中部分转写文本的字符的长度与对应的模板文本的字符的长度，选取长度长的作为字符总数，若轨迹矩阵中一轨迹中部分转写文本的字符的长度大于对应的模板文本的字符的长度，则在轨迹矩阵中一轨迹中选取部分转写文本的字符的长度为字符总数。若轨迹矩阵中一轨迹中部分转写文本的字符的长度小于对应的模板文本的字符的长度，则在轨迹矩阵中一轨迹中选取模板文本的字符的长度为字符总数。在选取长度长的作为字符总数之后，计算轨迹矩阵中各轨迹的相等字符个数与对应的字符总数的比值，在完成计算比值之后，获得轨迹矩阵中各轨迹的相似度。
- 20

获得模块 16，用于根据所述第一轨迹，确定所述部分转写文本在所述模板文本上对应的终点，获得第一终点。

- 25 在获得第一轨迹之后，根据第一轨迹，由于第一轨迹在轨迹矩阵中存在终点，确定部分转写文本在模板文本上对应的终点，从而获得第一终点。

在本实施例中，获得模块 16 包括：

第二标记模块，用于标记所述第一轨迹中最后一个元素；

- 30 第一获得模块，用于根据所述第一轨迹中最后一个元素，标记对应的所述模板文本的字符，获得第一终点。

在获得第一轨迹之后，标记第一轨迹中最后一个元素，根据第一轨迹中最后一个元素，获得第一轨

迹中最后一个元素对应列上的模板文本的字符，标记对应的模板文本的字符，从而获得第一终点。

第二获取模块 17，用于根据所述模板文本的始点和所述第一终点，从所述模板文本中获取新模板文本。

在获得第一终点之后，根据模板文本的始点和第一终点，获得两点之间的文本，包括模板文本的始点、第一终点分别对应的字符，从而从模板文本中获取新模板文本。

在本实施例中，第二获取模块 17 包括：

第三标记模块，用于在所述模板文本中标记所述模板文本的第一个字符为始点；

截取模块，用于截取所述模板文本的始点与所述第一终点之间的字符，其中所述模板文本的始点与所述第一终点之间的字符包括所述模板文本的始点对应的字符和所述第一终点对应的字符；

第二子获取模块，用于根据截取到的字符生成文本，获取所述新模板文本。

在得到第一终点之后，在模板文本中标记模板文本的第一个字符为始点，截取模板文本的始点与第一终点之间的字符，其中模板文本的始点与第一终点之间的字符包括模板文本的始点对应的字符和第一终点对应的字符。在截取到的字符之后，根据截取到的字符，生成与模板文本同样格式的文本，获得新模板文本。

第二计算模块 18，用于将所述部分转写文本与所述新模板文本进行对比，通过编辑距离算法计算所述部分转写文本的准确率。

在获得新模板文本之后，将部分转写文本与新模板文本进行对比，并不是与模板文本进行对比，通过编辑距离算法计算部分转写文本的准确率，从而解决现有的文本的转写准确率算法，将已经转写出来的文本与模板文本的全部文本进行对比，在部分文本被转写出来的时候，不能准确计算文本的转写准确率的问题。

如图 3 所示，本申请实施例中还提供一种计算机设备，该计算机设备可以是服务器，其内部结构可以如图 3 所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中，该计算机设计的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机程序和数据库。该内存储器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的数据库用于存储基于语义解析的文本准确率计算方法的模型等数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机程序被处理器执行时以实现一种基于语义解析的文本准确率计算方法。

上述处理器执行上述基于语义解析的文本准确率计算方法的步骤：获取从模板文本的始点开始被转写的部分转写文本；以所述模板文本字符的长度增加两个字符的长度为列数、以所述部分转写文本字符的长度增加两个字符的长度为行数，建立编辑距离矩阵；根据所述部分转写文本、所述模板文本，计算所述编辑距离矩阵中各元素的值；记录所述编辑距离矩阵中各元素的值的计算轨迹，生成与所述编辑距

离矩阵对应的轨迹矩阵；计算所述轨迹矩阵中各轨迹的相似度，筛选所述部分转写文本与所述模板文本相似度最高的一条轨迹，获得第一轨迹；根据所述第一轨迹，确定所述部分转写文本在所述模板文本上对应的终点，获得第一终点；根据所述模板文本的始点和所述第一终点，从所述模板文本中获取新模板文本；将所述部分转写文本与所述新模板文本进行对比，通过编辑距离算法计算所述部分转写文本的准确5 率。

本领域技术人员可以理解，图3中示出的结构，仅仅是与本发明方案相关的部分结构的框图，并不构成对本发明方案所应用于其上的计算机设备的限定。

本发明一实施例还提供一种计算机非易失性可读存储介质，其上存储有计算机程序，计算机程序被处理器执行时实现一种基于语义解析的文本准确率计算方法，具体为：获取从模板文本的始点开始被转10 写的部分转写文本；以所述模板文本字符的长度增加两个字符的长度为列数、以所述部分转写文本字符的长度增加两个字符的长度为行数，建立编辑距离矩阵；根据所述部分转写文本、所述模板文本，计算所述编辑距离矩阵中各元素的值；记录所述编辑距离矩阵中各元素的值的计算轨迹，生成与所述编辑距离矩阵对应的轨迹矩阵；计算所述轨迹矩阵中各轨迹的相似度，筛选所述部分转写文本与所述模板文本相似度最高的一条轨迹，获得第一轨迹；根据所述第一轨迹，确定所述部分转写文本在所述模板文本上15 对应的终点，获得第一终点；根据所述模板文本的始点和所述第一终点，从所述模板文本中获取新模板文本；将所述部分转写文本与所述新模板文本进行对比，通过编辑距离算法计算所述部分转写文本的准确率。

权利要求书

1、一种基于语义解析的文本准确率计算方法，其特征在于，所述方法包括：

5 获取从模板文本的始点开始被转写的部分转写文本；

以所述模板文本字符的长度增加两个字符的长度为列数、以所述部分转写文本字符的长度增加两个字符的长度为行数，建立编辑距离矩阵；

根据所述部分转写文本、所述模板文本，计算所述编辑距离矩阵中各元素的值；

记录所述编辑距离矩阵中各元素的值的计算轨迹，生成与所述编辑距离矩阵对应的轨迹矩阵；

10 计算所述轨迹矩阵中各轨迹的相似度，筛选所述部分转写文本与所述模板文本相似度最高的一条轨迹，获得第一轨迹；

根据所述第一轨迹，确定所述部分转写文本在所述模板文本上对应的终点，获得第一终点；

根据所述模板文本的始点和所述第一终点，从所述模板文本中获取新模板文本；

15 将所述部分转写文本与所述新模板文本进行对比，通过编辑距离算法计算所述部分转写文本的准确率。

2、根据权利要求 1 所述的基于语义解析的文本准确率计算方法，其特征在于，在所述以所述模板文本字符的长度增加两个字符的长度为列数、以所述部分转写文本字符的长度增加两个字符的长度为行数，建立编辑距离矩阵的步骤之后，在所述根据所述部分转写文本、所述模板文本，计算所述编辑距离矩阵中各元素的值的步骤之前，包括：

20 从所述编辑距离矩阵的第一行的第三个元素开始输入所述模板文本的字符；

从所述编辑距离矩阵的第一列的第三个元素开始输入所述部分转写文本的字符；

定义所述编辑距离矩阵的第二行中的第二个元素的值为 0；

以所述编辑距离矩阵的第二行中的第二个元素的值为 0 依次递增数值 1，初始化所述编辑距离矩阵的第二行的各元素的值；

25 以所述编辑距离矩阵的第二列中的第二个元素的值为 0 依次递增数值 1，初始化所述编辑距离矩阵的第二列的各元素的值。

3、根据权利要求 2 所述的基于语义解析的文本准确率计算方法，其特征在于，所述编辑距离矩阵中未被初始化的各元素的值由其左方、左上角、上方中的某一个元素的值来确定，在所述根据所述部分转写文本、所述模板文本，计算所述编辑距离矩阵中各元素的值的步骤中，包括：

30 识别所述编辑距离矩阵的第三列中的第三个元素所处的列数、行数；

识别所述编辑距离矩阵的第三列中的第三个元素所处的列数、行数分别对应所述模板文本的字符、所述部分转写文本的字符；

判断所述编辑距离矩阵的第三列中的第三个元素所处的列数对应所述模板文本的字符与所述编辑距离矩阵的第三列中的第三个元素处于行数对应所述部分转写文本的字符是否相等；

若所述编辑距离矩阵的第三列中的第三个元素所处的列数对应所述模板文本的字符与所述编辑距离矩阵的第三列中的第三个元素处于行数对应所述部分转写文本的字符相等，则所述编辑距离矩阵的第三列中的第三个元素的值为其左上角的元素的值；

若所述编辑距离矩阵的第三列中的第三个元素所处的列数对应所述模板文本的字符与所述编辑距离矩阵的第三列中的第三个元素处于行数对应所述部分转写文本的字符不相等，则所述编辑距离矩阵的第三列中的第三个元素的值为其左方、左上角、上方的元素中最小值加 1 得到；

依次计算所述编辑距离矩阵的第三列中的第四个元素的值，直至完成计算所述编辑距离矩阵中各元素的值。

4、根据权利要求 1 所述的基于语义解析的文本准确率计算方法，其特征在于，在所述记录所述编辑距离矩阵中各元素的值的计算轨迹，生成与所述编辑距离矩阵对应的轨迹矩阵的步骤中，包括：

记录所述编辑距离矩阵中各元素的值的计算轨迹；

根据所述编辑距离矩阵中各元素的值的计算轨迹，标记所述编辑距离矩阵中各元素的值产生由来；

在完成标记后，生成与所述编辑距离矩阵对应的轨迹矩阵。

5、根据权利要求 1 所述的基于语义解析的文本准确率计算方法，其特征在于，在所述计算所述轨迹矩阵中各轨迹的相似度的步骤中，包括：

识别所述轨迹矩阵中各轨迹中所述部分转写文本的字符与对应的所述模板文本的字符相等的个数，获得相等字符个数；

比较所述轨迹矩阵中各轨迹中所述部分转写文本的字符的长度与对应的所述模板文本的字符的长度，选取长度长的作为字符总数；

计算所述轨迹矩阵中各轨迹的相等字符个数与对应的字符总数的比值，获得所述轨迹矩阵中各轨迹的相似度。

6、根据权利要求 1 所述的基于语义解析的文本准确率计算方法，其特征在于，在所述根据所述第一轨迹，确定所述部分转写文本在所述模板文本上对应的终点，获得第一终点的步骤中，包括：

标记所述第一轨迹中最后一个元素；

根据所述第一轨迹中最后一个元素，标记对应的所述模板文本的字符，获得第一终点。

7、根据权利要求 6 所述的基于语义解析的文本准确率计算方法，其特征在于，在所述根据所述模板文本的始点和所述第一终点，从所述模板文本中获取新模板文本的步骤中，包括：

在所述模板文本中标记所述模板文本的第一个字符为始点；

截取所述模板文本的始点与所述第一终点之间的字符，其中所述模板文本的始点与所述第一终点之间的字符包括所述模板文本的始点对应的字符和所述第一终点对应的字符；

根据截取到的字符生成文本，获得所述新模板文本。

8、一种基于语义解析的文本准确率计算装置，其特征在于，所述装置包括：

第一获取模块，用于获取从模板文本的始点开始被转写的部分转写文本；

5 建立模块，用于以所述模板文本字符的长度增加两个字符的长度为列数、以所述部分转写文本字符的长度增加两个字符的长度为行数，建立编辑距离矩阵；

第一计算模块，用于根据所述部分转写文本、所述模板文本，计算所述编辑距离矩阵中各元素的值；

生成模块，用于记录所述编辑距离矩阵中各元素的值的计算轨迹，生成与所述编辑距离矩阵对应的轨迹矩阵；

10 筛选模块，用于计算所述轨迹矩阵中各轨迹的相似度，筛选所述部分转写文本与所述模板文本相似度最高的一条轨迹，获得第一轨迹；

获得模块，用于根据所述第一轨迹，确定所述部分转写文本在所述模板文本上对应的终点，获得第一终点；

第二获取模块，用于根据所述模板文本的始点和所述第一终点，从所述模板文本中获取新模板文本；

15 第二计算模块，用于将所述部分转写文本与所述新模板文本进行对比，通过编辑距离算法计算所述部分转写文本的准确率。

9、根据权利要求8所述的基于语义解析的文本准确率计算装置，其特征在于，所述装置包括：

第一输入模块，用于从所述编辑距离矩阵的第一行的第三个元素开始输入所述模板文本的字符；

第二输入模块，用于从所述编辑距离矩阵的第一列的第三个元素开始输入所述部分转写文本的字符；

20 定义模块，用于定义所述编辑距离矩阵的第二行中的第二个元素的值为0；

第一初始化模块，用于以所述编辑距离矩阵的第二行中的第二个元素的值为0依次递增数值1，初始化所述编辑距离矩阵的第二行的各元素的值；

第二初始化模块，用于以所述编辑距离矩阵的第二列中的第二个元素的值为0依次递增数值1，初始化所述编辑距离矩阵的第二列的各元素的值。

25 10、根据权利要求9所述的基于语义解析的文本准确率计算装置，其特征在于，所述第一计算模块包括：

第一识别模块，用于识别所述编辑距离矩阵的第三列中的第三个元素所处的列数、行数；

第二识别模块，用于识别所述编辑距离矩阵的第三列中的第三个元素所处的列数、行数分别对应所述模板文本的字符、所述部分转写文本的字符；

30 第一判断模块，用于判断所述编辑距离矩阵的第三列中的第三个元素所处的列数对应所述模板文本的字符与所述编辑距离矩阵的第三列中的第三个元素所处于行数对应所述部分转写文本的字符是否相等；若所述编辑距离矩阵的第三列中的第三个元素所处的列数对应所述模板文本的字符与所述编辑

距离矩阵的第三列中的第三个元素所处于行数对应所述部分转写文本的字符相等，则所述编辑距离矩阵的第三列中的第三个元素的值为其左上角的元素的值；若所述编辑距离矩阵的第三列中的第三个元素所处于的列数对应所述模板文本的字符与所述编辑距离矩阵的第三列中的第三个元素所处于行数对应所述部分转写文本的字符不相等，则所述编辑距离矩阵的第三列中的第三个元素的值为其左方、左上角、
5 上方的元素中最小值加 1 得到；

第一子计算模块，用于依次计算所述编辑距离矩阵的第三列中的第四个元素的值，直至完成计算所述编辑距离矩阵中各元素的值。

11、根据权利要求 8 所述的基于语义解析的文本准确率计算装置，其特征在于，所述生成模块包括：

第一记录模块，用于记录所述编辑距离矩阵中各元素的值的计算轨迹；

10 第一标记模块，用于根据所述编辑距离矩阵中各元素的值的计算轨迹，标记所述编辑距离矩阵中各元素的值产生由来；

第一生成模块，用于在完成标记后，生成与所述编辑距离矩阵对应的轨迹矩阵。

12、根据权利要求 8 所述的基于语义解析的文本准确率计算装置，其特征在于，所述筛选模块包括：

15 第三识别模块，用于识别所述轨迹矩阵中各轨迹中所述部分转写文本的字符与对应的所述模板文本的字符相等的个数，获得相等字符个数；

第一比较模块，用于比较所述轨迹矩阵中各轨迹中所述部分转写文本的字符的长度与对应的所述模板文本的字符的长度，选取长度长的作为字符总数；

第三计算模块，用于计算所述轨迹矩阵中各轨迹的相等字符个数与对应的字符总数的比值，获得所述轨迹矩阵中各轨迹的相似度。

20 13、根据权利要求 8 所述的基于语义解析的文本准确率计算装置，其特征在于，所述获得模块包括：

第二标记模块，用于标记所述第一轨迹中最后一个元素；

第一获得模块，用于根据所述第一轨迹中最后一个元素，标记对应的所述模板文本的字符，获得第一终点。

25 14、根据权利要求 13 所述的基于语义解析的文本准确率计算装置，其特征在于，所述第二获取模块包括：

第三标记模块，用于在所述模板文本中标记所述模板文本的第一个字符为始点；

截取模块，用于截取模块，用于截取所述模板文本的始点与所述第一终点之间的字符，其中所述模板文本的始点与所述第一终点之间的字符包括所述模板文本的始点对应的字符和所述第一终点对应的字符；

30 第二子获取模块，用于根据截取到的字符生成文本，获得所述新模板文本。

15、一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，其特征在于，所述处理器执行所述计算机程序时实现基于语义解析的文本准确率计算方法，所述方法包括：

获取从模板文本的始点开始被转写的部分转写文本;

以所述模板文本字符的长度增加两个字符的长度为列数、以所述部分转写文本字符的长度增加两个字符的长度为行数,建立编辑距离矩阵;

根据所述部分转写文本、所述模板文本,计算所述编辑距离矩阵中各元素的值;

5 记录所述编辑距离矩阵中各元素的值的计算轨迹,生成与所述编辑距离矩阵对应的轨迹矩阵;

计算所述轨迹矩阵中各轨迹的相似度,筛选所述部分转写文本与所述模板文本相似度最高的一条轨迹,获得第一轨迹;

根据所述第一轨迹,确定所述部分转写文本在所述模板文本上对应的终点,获得第一终点;

根据所述模板文本的始点和所述第一终点,从所述模板文本中获取新模板文本;

10 将所述部分转写文本与所述新模板文本进行对比,通过编辑距离算法计算所述部分转写文本的准确率。

16、根据权利要求 15 所述的计算机设备,其特征在于,在所述以所述模板文本字符的长度增加两个字符的长度为列数、以所述部分转写文本字符的长度增加两个字符的长度为行数,建立编辑距离矩阵的步骤之后,在所述根据所述部分转写文本、所述模板文本,计算所述编辑距离矩阵中各元素的值的步骤之前,包括:

从所述编辑距离矩阵的第一行的第三个元素开始输入所述模板文本的字符;

从所述编辑距离矩阵的第一列的第三个元素开始输入所述部分转写文本的字符;

定义所述编辑距离矩阵的第二行中的第二个元素的值为 0;

15 以所述编辑距离矩阵的第二行中的第二个元素的值为 0 依次递增数值 1,初始化所述编辑距离矩阵的第二行的各元素的值;

以所述编辑距离矩阵的第二列中的第二个元素的值为 0 依次递增数值 1,初始化所述编辑距离矩阵的第二列的各元素的值。

17、根据权利要求 16 所述的计算机设备,其特征在于,所述编辑距离矩阵中未被初始化的各元素的值由其左方、左上角、上方中的某一个元素的值来确定,在所述根据所述部分转写文本、所述模板文本,计算所述编辑距离矩阵中各元素的值的步骤中,包括:

识别所述编辑距离矩阵的第三列中的第三个元素所处的列数、行数;

识别所述编辑距离矩阵的第三列中的第三个元素所处的列数、行数分别对应所述模板文本的字符、所述部分转写文本的字符;

20 判断所述编辑距离矩阵的第三列中的第三个元素所处的列数对应所述模板文本的字符与所述编辑距离矩阵的第三列中的第三个元素处于行数对应所述部分转写文本的字符是否相等;

若所述编辑距离矩阵的第三列中的第三个元素所处的列数对应所述模板文本的字符与所述编辑距离矩阵的第三列中的第三个元素处于行数对应所述部分转写文本的字符相等,则所述编辑距离矩阵

的第三列中的第三个元素的值为其左上角的元素的值;

若所述编辑距离矩阵的第三列中的第三个元素所处于的列数对应所述模板文本的字符与所述编辑距离矩阵的第三列中的第三个元素所处于行数对应所述部分转写文本的字符不相等,则所述编辑距离矩阵的第三列中的第三个元素的值为其左方、左上角、上方的元素中最小值加1得到;

- 5 依次计算所述编辑距离矩阵的第三列中的第四个元素的值,直至完成计算所述编辑距离矩阵中各元素的值。

18、根据权利要求15所述的计算机设备,其特征在于,在所述记录所述编辑距离矩阵中各元素的值的计算轨迹,生成与所述编辑距离矩阵对应的轨迹矩阵的步骤中,包括:

记录所述编辑距离矩阵中各元素的值的计算轨迹;

- 10 根据所述编辑距离矩阵中各元素的值的计算轨迹,标记所述编辑距离矩阵中各元素的值产生由来;在完成标记后,生成与所述编辑距离矩阵对应的轨迹矩阵。

19、一种计算机非易失性可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现基于语义解析的文本准确率计算方法,所述方法包括:

获取从模板文本的始点开始被转写的部分转写文本;

- 15 以所述模板文本字符的长度增加两个字符的长度为列数、以所述部分转写文本字符的长度增加两个字符的长度为行数,建立编辑距离矩阵;

根据所述部分转写文本、所述模板文本,计算所述编辑距离矩阵中各元素的值;

记录所述编辑距离矩阵中各元素的值的计算轨迹,生成与所述编辑距离矩阵对应的轨迹矩阵;

计算所述轨迹矩阵中各轨迹的相似度,筛选所述部分转写文本与所述模板文本相似度最高的一条轨

- 20 迹,获得第一轨迹;

根据所述第一轨迹,确定所述部分转写文本在所述模板文本上对应的终点,获得第一终点;

根据所述模板文本的始点和所述第一终点,从所述模板文本中获取新模板文本;

将所述部分转写文本与所述新模板文本进行对比,通过编辑距离算法计算所述部分转写文本的准确率。

- 25 20、根据权利要求19所述的计算机非易失性可读存储介质,其特征在于,在所述以所述模板文本字符的长度增加两个字符的长度为列数、以所述部分转写文本字符的长度增加两个字符的长度为行数,建立编辑距离矩阵的步骤之后,在所述根据所述部分转写文本、所述模板文本,计算所述编辑距离矩阵中各元素的值的步骤之前,包括:

从所述编辑距离矩阵的第一行的第三个元素开始输入所述模板文本的字符;

- 30 从所述编辑距离矩阵的第一列的第三个元素开始输入所述部分转写文本的字符;

定义所述编辑距离矩阵的第二行中的第二个元素的值为0;

以所述编辑距离矩阵的第二行中的第二个元素的值为0依次递增数值1,初始化所述编辑距离矩阵

的第二行的各元素的值;

以所述编辑距离矩阵的第二列中的第二个元素的值为 0 依次递增数值 1, 初始化所述编辑距离矩阵的第二列的各元素的值。

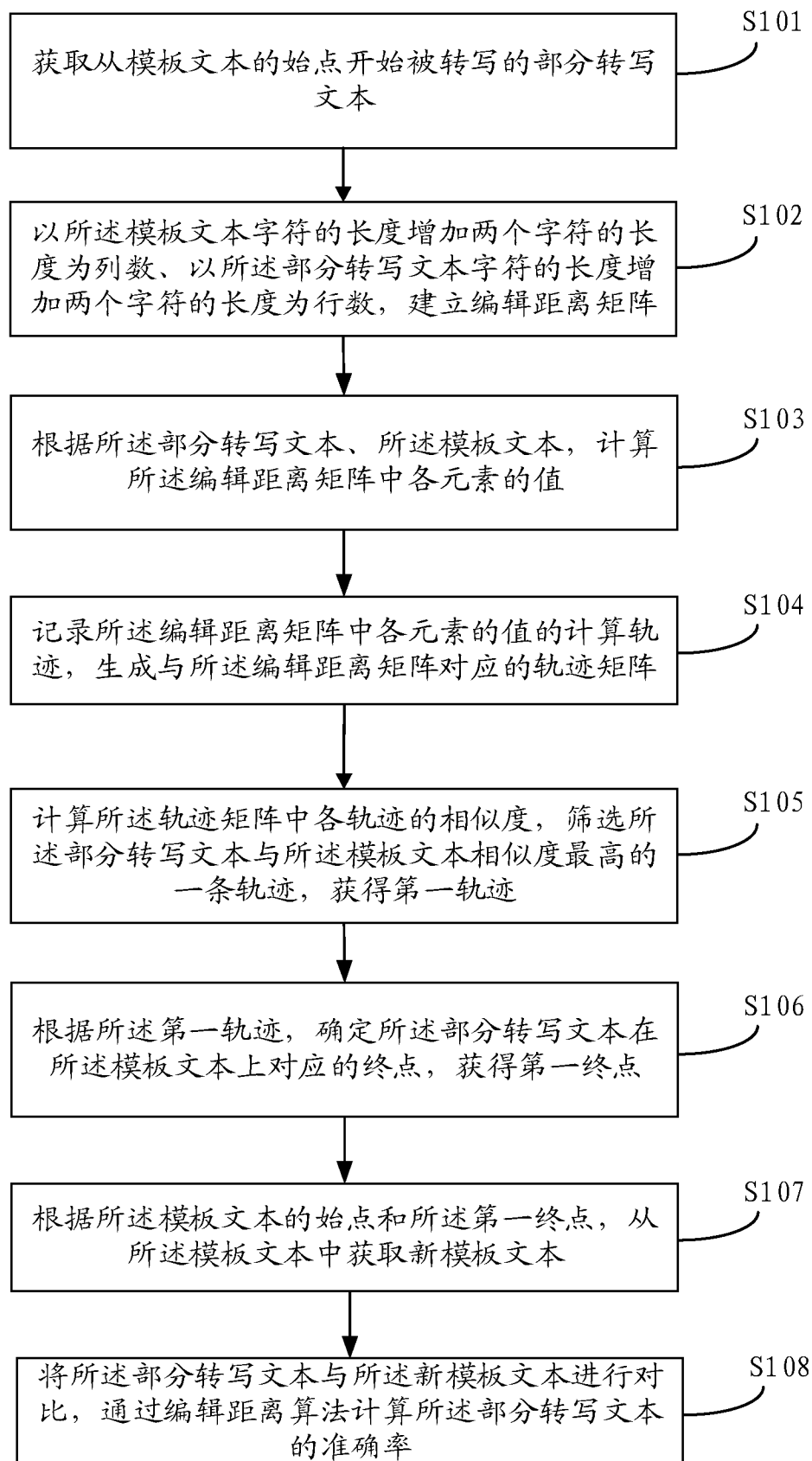


图 1

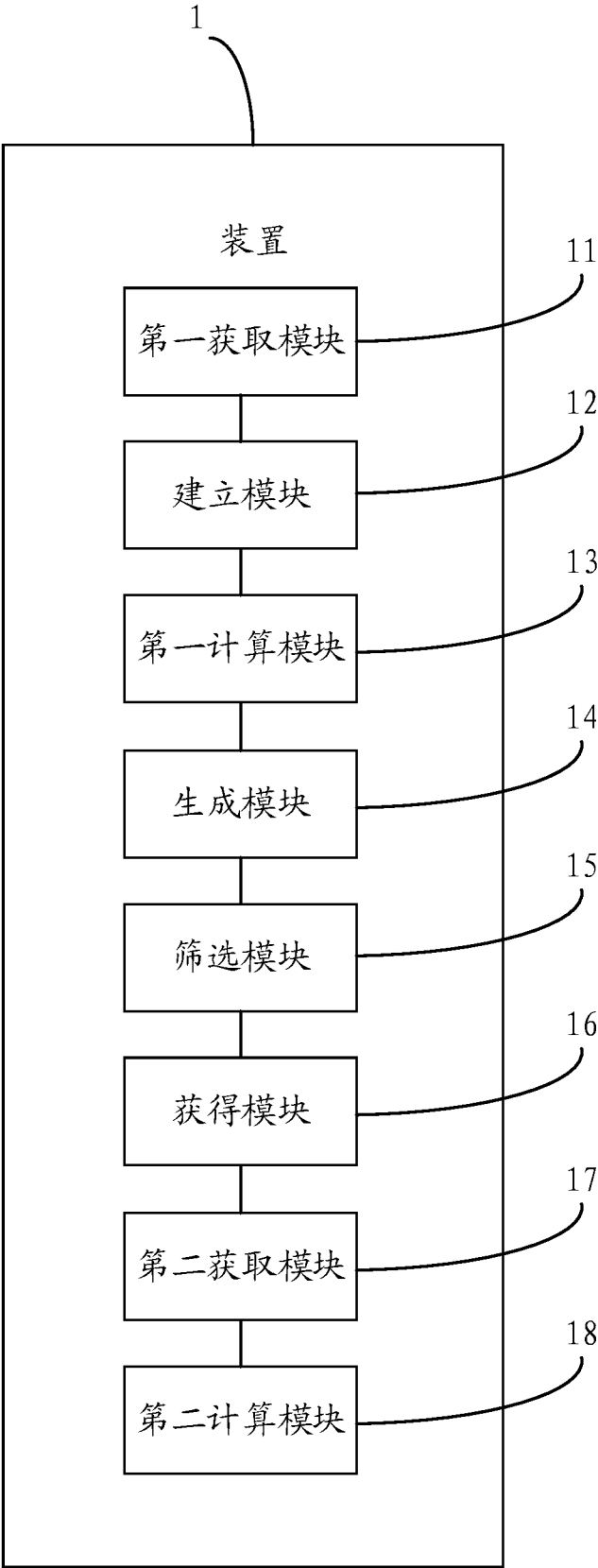


图 2

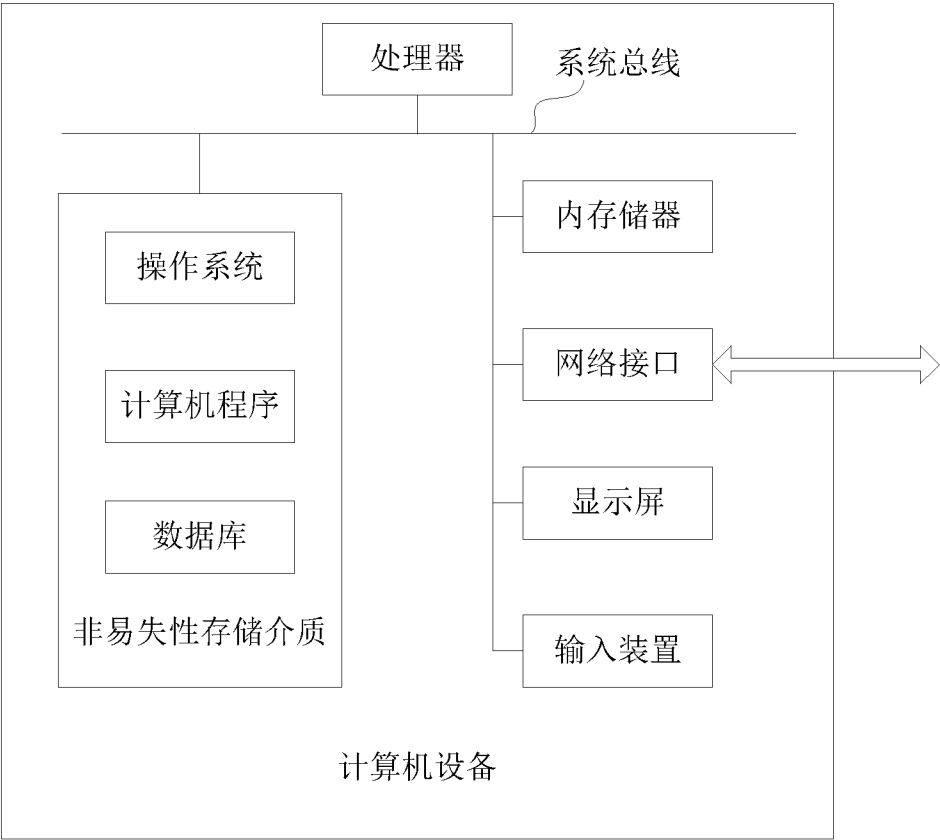


图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/124398

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/24(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT; CNABS; CNKI; USTXT; VEN; EPTXT; WOTXT: 转写, 开始, 结束, 终点, 文本, 矩阵, 轨迹, 编辑距离, 相似度, 模板, 对比, transcribe, start, end, text, matrix, trace, edit distance, similar, template, compare

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 103699591 A (HU'NAN UNIVERSITY) 02 April 2014 (2014-04-02) description, paragraphs [0024]-[0034], and figures 1-5	1-20
A	CN 108399163 A (BEIJING INSTITUTE OF TECHNOLOGY) 14 August 2018 (2018-08-14) description, paragraph [0005]	1-20
A	US 8001136 B1 (GOOGLE INC.) 16 August 2011 (2011-08-16) entire document	1-20

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“D” document cited by the applicant in the international application

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

19 July 2019

Date of mailing of the international search report

31 July 2019

Name and mailing address of the ISA/CN

**China National Intellectual Property Administration (ISA/
CN)**
**No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
 100088**
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/124398

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	103699591	A	02 April 2014	None	
CN	108399163	A	14 August 2018	None	
US	8001136	B1	16 August 2011	None	

国际检索报告

国际申请号

PCT/CN2018/124398

A. 主题的分类

G06F 17/24 (2006.01) i

按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献 (标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用))

CNTXT; CNABS; CNKI; USTXT; VEN; EPTXT; WOTXT; 转写, 开始, 结束, 终点, 文本, 矩阵, 轨迹, 编辑距离, 相似度, 模板, 对比, transcribe, start, end, text, matrix, trace, edit distance, similar, template, compare

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	CN 103699591 A (湖南大学) 2014年 4月 2日 (2014 - 04 - 02) 说明书第24-34段, 图1-5	1-20
A	CN 108399163 A (北京理工大学) 2018年 8月 14日 (2018 - 08 - 14) 说明书第5段	1-20
A	US 8001136 B1 (GOOGLE INC) 2011年 8月 16日 (2011 - 08 - 16) 全文	1-20

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2019年 7月 19日

国际检索报告邮寄日期

2019年 7月 31日

ISA/CN的名称和邮寄地址

中国国家知识产权局 (ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

授权官员

刘细金

电话号码 86-(20)-28950395

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/124398

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	103699591	A	2014年 4月 2日	无	
CN	108399163	A	2018年 8月 14日	无	
US	8001136	B1	2011年 8月 16日	无	