(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

(43) 国际公布日
2019 年 7 月 18 日 (18.07.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/136993 A1

(51) 国际专利分类号:
G06F 17/27 (2006.01)

(21) 国际申请号: PCT/CN2018/099994

(22) 国际申请日: 2018 年 8 月 10 日 (10.08.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810031770.0 2018年1月12日 (12.01.2018) CN

(71) 申请人: 深圳壹账通智能科技有限公司(**ONE CONNECT SMART TECHNOLOGY CO., LTD. (SHENZHEN)**) [CN/CN]; 中国广东省深圳市前海深港合作区前湾一路 1 号 A 栋 201 室, Guangdong 518052 (CN)。

(72) 发明人: 艾明 (**AI, Ming**); 中国广东省深圳市前海深港合作区前湾一路 1 号 A 栋 201 室, Guangdong 518052 (CN)。

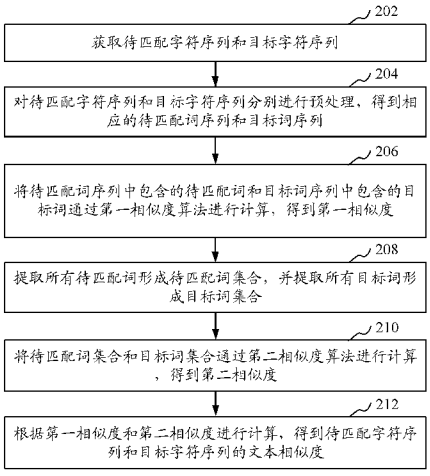
(74) 代理人: 广州华进联合专利商标代理有限公司 (**ADVANCE CHINA IP LAW OFFICE**); 中国广东省广州市天河区珠江东路 6 号 4501 房（部位：自编 01-03 和 08-12 单元）（仅限办公用途）, Guangdong 510623 (CN)。

(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, 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,

(54) **Title:** TEXT SIMILARITY CALCULATION METHOD AND DEVICE, COMPUTER APPARATUS, AND STORAGE MEDIUM

(54) 发明名称: 文本相似度计算方法、装置、计算机设备和存储介质



202 OBTAIN A CHARACTER SEQUENCE TO BE MATCHED AND A TARGET CHARACTER SEQUENCE
204 PREPROCESS THE CHARACTER SEQUENCE TO BE MATCHED AND THE TARGET CHARACTER SEQUENCE TO OBTAIN A CORRESPONDING WORD SEQUENCE TO BE MATCHED AND A CORRESPONDING TARGET WORD SEQUENCE
206 PERFORM CALCULATION, BY MEANS OF A FIRST SIMILARITY ALGORITHM, ON A WORD TO BE MATCHED CONTAINED IN THE WORD SEQUENCE TO BE MATCHED AND A TARGET WORD CONTAINED IN THE TARGET WORD SEQUENCE, SO AS TO OBTAIN A FIRST SIMILARITY DEGREE
208 EXTRACT ALL OF THE WORDS TO BE MATCHED, SO AS TO FORM A SET OF WORDS TO BE MATCHED, AND EXTRACT ALL OF THE TARGET WORDS TO FORM A TARGET WORD SET
210 PERFORM CALCULATION ON THE SET OF WORDS TO BE MATCHED AND THE TARGET WORD SET BY MEANS OF A SECOND SIMILARITY ALGORITHM, SO AS TO OBTAIN A SECOND SIMILARITY DEGREE
212 CALCULATE, ACCORDING TO THE FIRST SIMILARITY DEGREE AND THE SECOND SIMILARITY DEGREE, A TEXT SIMILARITY DEGREE BETWEEN THE CHARACTER SEQUENCE TO BE MATCHED AND THE TARGET CHARACTER SEQUENCE

图 2

(57) **Abstract:** A text similarity calculation method comprises: obtaining a character sequence to be matched and a target character sequence; respectively preprocessing the character sequence to be matched and the target character sequence to obtain a corresponding word sequence to be matched and a corresponding target word sequence; performing calculation, by means of a first similarity algorithm, on a word to be matched contained in the word sequence to be matched and a target word contained in the target word sequence, so as to obtain a first similarity degree; extracting all of the words to be matched, so as to form a set of words to be matched, and extracting all of the target words to form a target word set; performing calculation on the set of words to be matched and the target word set by means of a second similarity algorithm, so as to obtain a second similarity degree; and calculating, according to the first similarity degree and the second similarity degree, a text similarity degree between the character sequence to be matched and the target character sequence.

WO 2019/136993 A1

[见续页]

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: 一种文本相似度计算方法，包括：获取待匹配字符序列和目标字符序列；对待匹配字符序列和目标字符序列分别进行预处理，得到相应的待匹配词序列和目标词序列；将待匹配词序列中包含的待匹配词和目标词序列中包含的目标词通过第一相似度算法进行计算，得到第一相似度；提取所有待匹配词形成待匹配词集合，并提取所有目标词形成目标词集合；将待匹配词集合和目标词集合通过第二相似度算法进行计算，得到第二相似度；及根据第一相似度和第二相似度进行计算，得到待匹配字符序列和目标字符序列的文本相似度。

文本相似度计算方法、装置、计算机设备和存储介质

本申请要求于 2018 年 01 月 12 日提交中国专利局，申请号为 2018100317700，申请名称为“文本相似度计算方法、装置、计算机设备和存储介质”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。

技术领域

本申请涉及一种文本相似度计算方法、装置、计算机设备和存储介质。

背景技术

随着聊天机器人技术的发展，出现了字符串模糊搜索的概念，通常会采取编辑距离算法实现字符串匹配。编辑距离又称 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。许可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。编辑距离值越大，则字符串之间的相似度越小。

然而，由于语言的复杂性，相同的意思可以通过不同的文本来进行表达，而表面上很相似的文本，所表达出的含义也可能大不相同，传统的编辑距离算法通常以单个字符为单位计算各个字符序列之间的编辑距离，计算出来的编辑距离只是文字表面的距离，导致计算得到的文本相似度的准确度较低。

发明内容

根据本申请公开的各种实施例，提供一种能够提高文本相似度的文本相似度计算方法、装置、计算机设备和存储介质。

一种文本相似度计算方法，包括：获取待匹配字符序列和目标字符序列；对所述待匹配字符序列和所述目标字符序列分别进行预处理，得到相应的待匹配词序列和目标词序列； 将所述待匹配词序列中包含的待匹配词和所述目标词序列中包含的目标词通过第一相似度算法进行计算，得到第一相似度；提取所有待匹配词形成待匹配词集合，并提取所有目标词形成目标词集合；将所述待匹配词集合和所述目标词集合通过第二相似度算法进行计算，得到第二相似度；及根据所述第一相似度和所述第二相似度进行计算，得到所述待匹配字符序列和所述目标字符序列的文本相似度。

一种文本相似度计算装置，包括：字符序列获取模块，用于获取待匹配字符序列和目标字符序列；词序列获取模块，用于对所述待匹配字符序列和所述目标字符序列分别进行预处理，得到相应的待匹配词序列和目标词序列； 第一相似度计算模块，用于将所述待匹配词序列中包含的待匹配词和所述目标词序列中包含的目标词通过第一相似度算法进行计算，得到第一相似度；词集合形成模块，用于提取所有待匹配词形成待匹配词集合，

并提取所有目标词形成目标词集合；第二相似度计算模块，用于将所述待匹配词集合和所述目标词集合通过第二相似度算法进行计算，得到第二相似度；及文本相似度计算模块，用于根据所述第一相似度和所述第二相似度进行计算，得到所述待匹配字符序列和所述目标字符序列的文本相似度。

一种计算机设备，包括存储器和一个或多个处理器，所述存储器中储存有计算机可读指令，所述计算机可读指令被所述处理器执行时，使得所述一个或多个处理器执行以下步骤：获取待匹配字符序列和目标字符序列；对所述待匹配字符序列和所述目标字符序列分别进行预处理，得到相应的待匹配词序列和目标词序列； 将所述待匹配词序列中包含的待匹配词和所述目标词序列中包含的目标词通过第一相似度算法进行计算，得到第一相似度；提取所有待匹配词形成待匹配词集合，并提取所有目标词形成目标词集合；将所述待匹配词集合和所述目标词集合通过第二相似度算法进行计算，得到第二相似度；及根据所述第一相似度和所述第二相似度进行计算，得到所述待匹配字符序列和所述目标字符序列的文本相似度。

一个或多个存储有计算机可读指令的非易失性计算机可读存储介质，计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行以下步骤：获取待匹配字符序列和目标字符序列；对所述待匹配字符序列和所述目标字符序列分别进行预处理，得到相应的待匹配词序列和目标词序列； 将所述待匹配词序列中包含的待匹配词和所述目标词序列中包含的目标词通过第一相似度算法进行计算，得到第一相似度；提取所有待匹配词形成待匹配词集合，并提取所有目标词形成目标词集合；将所述待匹配词集合和所述目标词集合通过第二相似度算法进行计算，得到第二相似度；及根据所述第一相似度和所述第二相似度进行计算，得到所述待匹配字符序列和所述目标字符序列的文本相似度。

本申请的一个或多个实施例的细节在下面的附图和描述中提出。本申请的其它特征和优点将从说明书、附图以及权利要求书变得明显。

附图说明

为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1 为根据一个或多个实施例中文本相似度计算方法的应用场景图。

图 2 为根据一个或多个实施例中文本相似度计算方法的流程示意图。

图 3A 为根据一个或多个实施例中实体物质衍生的词语树的示意图。

图 3B 为根据一个或多个实施例中虚拟事件衍生的词语树的示意图。

图 4 为根据另一个或多个实施例中文本相似度计算方法的流程示意图。

图 5 为根据一个或多个实施例中文本相似度计算装置的结构框图。

图 6 为根据一个或多个实施例中计算机设备的内部结构图。

具体实施方式

为了使本申请的技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

可以理解，本申请所使用的术语“第一”、“第二”等可在本文中用于描述各种元件，但这些元件不受这些术语的限制。这些术语仅用于将第一个元件与另一个元件区分。举例来说，在不脱离本申请的范围的情况下，可以将第一相似度称为第二相似度，且类似地，可将第二相似度称为第一相似度。第一相似度和第二相似度两者都是相似度，但其不是同一相似度。

本申请提供的文本相似度计算方法，可以应用于如图 1 所示的应用环境中。终端 102 通过网络与服务器 104 通过网络进行通信。比如说服务器 104 可接收终端 102 发送的待匹配字符序列。终端 102 可以但不限于是各种个人计算机、笔记本电脑、智能手机、平板电脑和便携式可穿戴设备，服务器 104 可以用独立的服务器或者是多个服务器组成的服务器集群来实现。

在其中一个实施例中，如图 2 所示，提供了一种文本相似度计算方法，以该方法应用于图 1 中的服务器 104 为例进行说明，包括以下步骤:

步骤 202，获取待匹配字符序列和目标字符序列。

待匹配字符序列是指需要进行匹配的字符序列。目标字符序列是指数据库中预设的字符序列，用于与待匹配字符序列进行匹配。字符序列是指以字符为单位按顺序形成的序列，字符可以是字母、阿拉伯数字、汉字和标点符号中的至少一种。字符序列包括但不限于字母、阿拉伯数字、汉字和标点符号等其中一种或多种的组合。

步骤 204，对待匹配字符序列和目标字符序列分别进行预处理，得到相应的待匹配词序列和目标词序列。

预处理是指对待匹配字符序列和目标字符序列进行转换、缩减、拆分等其中至少一种操作的过程。待匹配词序列是指由待匹配字符序列经过预处理得到的词序列。目标词序列是指由目标字符序列经过预处理得到的词序列。词序列是指以词语为单位按顺序形成的序列。待匹配词序列是指以待匹配词为单位按顺序形成的序列。目标词序列是指以目标词为单位按顺序形成的序列。待匹配词和目标词可为由一个或多个字符构成的单纯词，也可为由两个及以上的单纯词构成的合成词。

在其中一个实施例中，步骤 204 包括：将待匹配字符序列中包含的无关字符和目标字符序列中包含的无关字符删除；对删除无关字符后的待匹配字符序列和目标字符序列分别进行分词，得到相应的待匹配词序列和目标词序列。

无关字符是指不影响文本相似度计算的字符，包括但不限于标点符号及停用字符。

分词是指将字符序列按照一定的规律转换成词序列的处理过程。可采用基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法等其中一种或多种的分词方法，对删除无关字符后的待匹配字符序列和目标字符序列分别进行分词。

步骤206，将待匹配词序列中包含的待匹配词和目标词序列中包含的目标词通过第一相似度算法进行计算，得到第一相似度。

第一相似度算法是指按照两个词序列中词语的顺序进行逐词比较之后计算得到相似度的算法。将待匹配词序列与目标词序列分别作为一维词序列，按照待匹配词的顺序和目标词的顺序通过第一相似度算法进行计算，得到第一相似度。将待匹配词序列与目标词序列以一维的形式进行相似度计算，能够节约系统的存储空间，降低时间复杂度。

在其中一个实施例中，步骤206包括：将待匹配词序列中包含的待匹配词和目标词序列中包含的目标词通过编辑距离公式进行计算，得到待匹配词序列与目标词序列之间的编辑距离；获取待匹配词序列中包含的待匹配词的第一数量，和目标词序列中包含的目标词的第二数量；根据编辑距离、第一数量和第二数量进行计算，得到第一相似度。

编辑距离是指两个词序列之间，由一个转成另一个所需的最少编辑操作次数。以词语为单位计算两个词序列之间的编辑距离，能够降低词序列的语义对词序列编辑的影响，提高计算词序列相似度的准确性。

举例来说，长度为$|S|$的待匹配词序列 S 中包含$|S|$个待匹配词，长度为|T|的目标词序列 T 中包含$|T|$个目标词。待匹配词序列 S 与目标词序列 T 的编辑距离$lev_{S,T}(i,j)$可通过公式

$$lev_{S,T}(i,j)=\begin{cases} i & ,j=0 \\ j & ,i=0 \\ \min\begin{cases} lev_{S,T}(i,j-1)+1 \\ lev_{S,T}(i-1,j)+1 \\ lev_{S,T}(i-1,j-1)+1 \end{cases} & ,otherwise \end{cases}$$

计算得到。i表示待匹配词序列 S 中第i个待匹配词，j表示目标词序列 T 中第j个目标词。当i和j中存在至少一个为 0 时，编辑距离$lev_{S,T}(i,j)$取i和j中的最大值；否则，编辑距离$lev_{S,T}(i,j)$取$lev_{S,T}(i,j-1)+1$、$lev_{S,T}(i-1,j)+1$、$lev_{S,T}(i-1,j-1)+1$中的最小值。基于计算得到的编辑距离，可通过下列相似度计算公式$sim1_{S,T}(i,j)=\frac{Max(|S|,|T|)-lev_{S,T}(i,j)}{Max(|S|,|T|)}$计算得到第一相似度$sim1_{S,T}(i,j)$。$Max(|S|,|T|)$表示$|S|$、$|T|$中的最大值。且第一相似度$sim1_{S,T}(i,j)$取值为 0 至 1。

步骤208，提取所有待匹配词形成待匹配词集合，并提取所有目标词形成目标词集合。

待匹配词集合是指由待匹配词序列中包含的所有待匹配词构成的集合。待目标词集合是指由目标词序列中包含的所有目标词构成的集合。待匹配词集合中的待匹配词不具有

顺序，相应地，目标词集合中的目标词也不具有顺序。

在其中一个实施例中，还可将文字表达的数字转换为阿拉伯数字，比如说可将“三十三”转换为“33”。统一为阿拉伯数字，可以更快捷地进行数字的匹配，提高文本相似度的准确性。

步骤210，将待匹配词集合和目标词集合通过第二相似度算法进行计算，得到第二相似度。

第其中二相似度算法是将所有的待匹配词和所有的目标词分别作为一个整体进行比较的相似度算法。包括但不限于基于语义词典的词汇相似度算法及基于语料统计的词汇相似度算法等。

在其中一个实施例中，步骤210，包括：将待匹配词集合和目标词集合进行匹配，统计待匹配词与目标词的匹配数量；统计待匹配词集合的待匹配词数量和目标词集合的目标词数量；根据匹配数量、待匹配词数量和目标词数量进行计算，得到第二相似度。

步骤212，根据第一相似度和第二相似度进行计算，得到待匹配字符序列和目标字符序列的文本相似度。

文本相似度是指待匹配字符序列和目标字符序列之间的相似度。计算得到第一相似度和第二相似度之后，可将第一相似度乘以第二相似度作为文本相似度。还可预设与第一相似度的对应的第一权值，与第二相似度对应的第二权值，对第一相似度与第二相似度进行加权求和计算，得到文本相似度。

上述文本相似度计算方法中，在获取待匹配字符序列和目标字符序列之后，通过对匹配字符序列和目标字符序列进行预处理，得到以词语为单位按顺序形成的待匹配词序列和目标词序列，通过考虑词语顺序的第一相似度算法计算得到第一相似度，再根据待匹配词序列中包含的待匹配词和目标词序列中包含的目标词分别形成待匹配词集合与目标词集合，通过不考虑词语顺序的第二相似度算法计算得到第二相似度，然后综合第一相似度和第二相似度计算得到待匹配字符序列和目标字符序列之间的文本相似度。通过以词语为单位进行相似度计算，并综合两种相似度算法计算文本相似度，降低了以单字符通过单一相似度算法导致的误差，提高了文本相似度计算的准确性。

在其中一个实施例中，无关字符包括停用字符和相同字符；将待匹配字符序列中包含的无关字符和目标字符序列中包含的无关字符删除，包括：将待匹配字符序列中包含的停用字符和目标字符序列中包含的停用字符删除；判断删除停用字符后的待匹配字符序列和目标字符序列中是否存在相同字符；相同字符是指在删除停用字符后的待匹配字符序列和目标字符序列中，处于相同位置的相同字符；若是，则将删除停用字符后的待匹配字符序列和目标字符序列中包含的相同字符删除，得到相应的待匹配词序列和目标词序列。

停用字符是指在信息检索中，为节省存储空间和提高搜索效率，在处理字符序列之前可滤掉某些字或词。针对过滤停用字符可预设停用字符库。比如中文停用字符包括但不限于语气词、连接词及转折词等，比如“啊”、“吧”、“哎”、“的”、“此外”、“但是”等。

当检测到停用字符时，可将待匹配字符序列或目标字符序列中包含的停用字符删除。

由于待匹配字符序列中包含的待匹配字符和目标字符序列中包含的目标字符都具有顺序，可将待匹配字符和目标字符按顺序进行匹配，将在待匹配字符序列和目标字符序列中处于相同位置上的相同的字符作为相同字符。分别将待匹配字符序列中的相同字符与目标字符序列中的目标字符删除。举例来说，待匹配字符序列为“这个算法该如何优化”，目标字符序列为“优化算法该怎么做”，通过匹配可知，“算”和“法”处于待匹配字符序列和目标字符序列中的相同位置，因此，可以删除“算”和“法”。删除相同字符之后的待匹配字符序列为“这个该如何优化”，目标字符序列为“优化该怎么做”。

上述实施例中，通过删除停用字符和相同字符等无关字符，可以缩减参与文本相似度计算的词序列长度，能够节约文本相似度计算时间，减少计算所占用的内存空间，提高文本相似度计算效率。

在其中一个实施例中，还可以将待匹配字符序列或目标字符序列中包含的无关字符替换为预设字符，替换之后可将待匹配字符序列或目标字符序列中包含的预设字符全部清除。比如可将待匹配字符序列 S1 中包含的无关字符替换为空格符之后，得到包含空格符的待匹配字符序列 S2，将待匹配字符序列 S2 中包含的空格符全部清除，得到不包含空格符的待匹配字符序列 S3。可对待匹配字符序列 S3 进行分词，得到待匹配词序列 S4。

在其中一个实施例中，通过基于语义词典的词汇相似度算法计算第二相似度之前，可针对语义词典中词语的上下位层次关系可构建词语树，如图 3A 和图 3B 所示，图 3A 中的词语为实体物质衍生的词语树，图 3B 中的词语为虚拟事件衍生的词语树。父节点所对应的词语与子节点所对应的词语具有上下位的关系。可根据词语树计算词语之间的语义距离，且层次越高路径参数越大，层次越低，路径参数越小。距离越大，相似度越小。根据词语树计算词语 A 和词语 B 在词语树中的路径长度，即语义距离为 d 之后，可根据公式计算词语 A 和词语 B 的相似度 $sim(A,B)=\frac{\alpha}{d+\alpha}$，$\alpha$ 为参数。

在其中一个实施例中，可将待匹配词集合与目标词集合进行匹配，通过第二相似度算法计算得到待匹配词集合中的每个待匹配词与目标词集合中每个目标词的第二子相似度。根据计算得到的所有的第二子相似度可计算得到第二相似度。

在其中一个实施例中，还可将在提取所有待匹配词形成待匹配词集合，并提取所有目标词形成目标词集合之后，统计大于预设子相似度阈值的第二子相似度所对应的待匹配词的匹配数量 $Q(S,T)$，并统计待匹配词集合中包含的待匹配词数量 $|S|$ 和目标词集合中包含的目标词数量 $|T|$。第二相似度 $sim2$ 可通过公式 $sim2=\frac{Q(S,T)}{Max(|S|,|T|)}$ 计算得到。其中 $Max(|S|,|T|)$ 表示待匹配词数量 $|S|$ 和目标词数量 $|T|$ 中的最大值。

在其中一个实施例中，在获取待匹配字符序列和目标字符序列之后，还包括：获取待匹配字符序列对应的待匹配拼音序列和目标字符序列对应的目标拼音序列；将待匹配拼音序列中包含的待匹配拼音和目标拼音序列中包含的目标拼音通过第一相似度算法进行计算，得到第三相似度；根据第一相似度和第二相似度进行计算，得到待匹配字符序列和目标字符序列的文本相似度，包括：根据第一相似度、第二相似度和第三相似度进行计算，得到待匹配字符序列和目标字符序列的文本相似度。

待匹配拼音序列是指待匹配字符序列中的待匹配字符所对应的拼音构成的序列。目标拼音序列是指目标字符序列中的目标字符所对应的拼音构成的序列。可通过在用户进行输入操作时，获取用户输入的待匹配字符所对应的拼音，生成待匹配拼音序列。目标拼音序列可为数据库中预设的与目标字符序列所对应的序列。可以将待匹配拼音序列与目标字符序列以每个字符所对应的拼音为单位，通过第一相似度算法进行计算得到第三相似度。

举例来说，待匹配字符序列“你名字拗口”所对应的待匹配拼音序列为“ni ming zi ao kou”，目标字符序列“你太执拗”所对应的目标拼音序列为“ni tai zhi niu”。虽然待匹配字符序列和目标字符序列中都包含“拗”这个字符，但是由于“拗”在待匹配拼音序列和目标拼音序列中所对应的拼音分别为“ao”和“niu”，差别很大，因此通过以待匹配拼音序列和目标拼音序列进行文本相似度的计算，可以降低“拗”这个多音字带来的误差。

上述实施例中，通过引入待匹配拼音序列和目标拼音序列，能够检测到由于多音字导致相同文字但语义不同的情况，从而降低因多音字导致的文本相似度误差。

在其中一个实施例中，获取待匹配字符序列和目标字符序列，包括：接收终端发送的待匹配字符序列；根据待匹配字符序列从数据库中获取多个目标字符序列；在根据第一相似度和第二相似度进行计算，得到待匹配字符序列和目标字符序列的文本相似度之后，还包括：查询文本相似度大于预设相似度阈值的目标字符序列所对应的相关资源；将相关资源发送至终端。

将待匹配字符序列与多个目标字符序列进行文本相似度计算，还可以确定与待匹配字符序列的文本相似度最高的目标字符序列。目标字符序列可关联文本、图片、链接、音频、视频等相关资源。举例来说，待匹配字符序列可为用户通过终端发送的用于咨询问题的字符序列。目标字符序列可为关联相应答案文本的字符序列。当确定了与待匹配字符序列文本相似度最高的目标字符序列之后，可将目标字符序列关联的相应答案文本的字符序列发送至终端。

在其中一个实施例中，如图 4 所示，提供了另一种文本相似度计算方法，该方法包括以下步骤：

步骤 402，获取待匹配字符序列和目标字符序列。

待匹配字符序列和目标字符序列可为字母、阿拉伯数字、汉字和标点符号等其中一种或多种的组合。

举例来说，待匹配字符序列可以为用户通过终端发送的用于咨询问题的字符序列。比

如待匹配字符序列可以是“请问 3 台电脑是多少钱？”。而目标字符序列可以为数据库中预存的问题模板的字符序列。比如目标字符序列可以是“3 台计算机价格？”。当接收到终端发送的待匹配字符序列之后，可查找数据库中预设的目标字符序列。

步骤 404，将待匹配字符序列中包含的无关字符和目标字符序列中包含的无关字符删除。

无关字符包括但不限于标点符号及停用字符。针对过滤停用字符可预设停用字符库。中文停用字符包括但不限于语气词、连接词及转折词等，比如“啊”、“吧”、“哎”、“的”、“此外”、“但是”等。当检测到停用字符时，可将待匹配字符序列或目标字符序列中包含的停用字符删除。

举例来说，待匹配字符序列为“请问 3 台电脑是多少钱？”。包含标点符号“？”，和停用字符“是”，将该待匹配字符序列中包含的无关字符进行删除之后得到“3 台电脑多少钱”。

步骤 406，对删除无关字符后的待匹配字符序列和目标字符序列分别进行分词，得到相应的待匹配词序列和目标词序列。

分词是指将字符序列按照一定的规律转换成词序列的处理过程。待匹配词序列是指以待匹配词为单位按顺序形成的序列。目标词序列是指以目标词为单位按顺序形成的序列。

举例来说，对该删除无关字符之后的待匹配字符序列进行分词，可得到待匹配词序列“3|台|电脑|多少|钱”，其中“|”表示词分隔符，用于区分待匹配词序列中不同的词语。该待匹配词序列中包含“3”、“台”、“电脑”、“多少”、“钱”五个待匹配词。

步骤 408，将待匹配词序列中包含的待匹配词和目标词序列中包含的目标词通过编辑距离公式进行计算，得到待匹配词序列与目标词序列之间的编辑距离。

编辑距离公式是指计算两个词序列之间以词语为单位由一个转成另一个所需的最少编辑操作次数的公式。最少编辑操作次数即为编辑距离。许可的编辑操作包括将一个词语替换成另一个词语，插入一个词语，删除一个词语。

举例来说，待匹配词序列为“3|台|电脑|多少|钱”，目标词序列为“3|台|计算机|价格”。将待匹配词序列转换为目标词序列需要进行 3 次操作，包括将“电脑”替换成“计算机”，将“多少”删除，将“钱”替换为“价格”。还可以预设同义词词库，由于“电脑”等同于“计算机”，因此可以不将“电脑”与“计算机”的替换过程不计入编辑距离。

步骤 410，获取待匹配词序列中包含的待匹配词的第一数量，和目标词序列中包含的目标词的第二数量。

步骤 412，根据编辑距离、第一数量和第二数量进行计算，得到第一相似度。

举例来说，长度为$|S|$的待匹配词序列 S 中包含$|S|$个待匹配词，长度为|T|的目标词序列 T 中包含$|T|$个目标词。待匹配词序列 S 与目标词序列 T 的编辑距离$lev_{S,T}(i,j)$可通过公

式 $lev_{S,T}(i,j)=\begin{cases} i & ,j=0 \\ j & ,i=0 \\ \min\begin{cases} lev_{S,T}(i,j-1)+1 \\ lev_{S,T}(i-1,j)+1 \\ lev_{S,T}(i-1,j-1)+1 \end{cases} & ,otherwise \end{cases}$ 计算得到。i 表示待匹配词序列 S 中第 i 个待匹配词，j 表示目标词序列 T 中第 j 个目标词。当 i 和 j 中存在至少一个为 0 时，编辑距离 $lev_{S,T}(i,j)$ 取 i 和 j 中的最大值；否则，编辑距离 $lev_{S,T}(i,j)$ 取 $lev_{S,T}(i,j-1)+1$、$lev_{S,T}(i-1,j)+1$、$lev_{S,T}(i-1,j-1)+1$ 中的最小值。基于计算得到的编辑距离，可通过下列相似度计算公式 $sim1_{S,T}(i,j)=\frac{Max(|S|,|T|)-lev_{S,T}(i,j)}{Max(|S|,|T|)}$ 计算得到第一相似度 $sim1_{S,T}(i,j)$。$Max(|S|,|T|)$ 表示 $|S|$、$|T|$ 中的最大值。且第一相似度 $sim1_{S,T}(i,j)$ 取值为 0 至 1。

步骤 414，提取所有待匹配词形成待匹配词集合，并提取所有目标词形成目标词集合。

举例来说，待匹配词序列为“3|台|电脑|多少|钱”，可提取待匹配词序列中包含的所有待匹配词，形成待匹配词集合为{“3”、“台”、“电脑”、“多少”、“钱”}。“3”、“台”、“电脑”、“多少”及“钱”五个待匹配词是并列关系，不具有顺序。

步骤 416，将待匹配词集合和目标词集合进行匹配，统计待匹配词与目标词的匹配数量。

可针对语义词典中词语的上下位层次关系可构建词语树，通过待匹配词与目标词在词语树中的路径距离，计算待匹配词与目标词之间的第二子相似度。将大于预设子相似度阈值的第二子相似度所对应的待匹配词与目标词判定为匹配，统计待匹配词集合和目标词集合中待匹配词与目标词的匹配数量。

步骤 418，统计待匹配词集合的待匹配词数量和目标词集合的目标词数量。

步骤 420，根据匹配数量、待匹配词数量和目标词数量进行计算，得到第二相似度。

举例来说，待匹配词集合为{“电脑”、“多少”、“钱”}，目标词集合为{“计算机”、“价格”}，可计算“电脑”与“计算机”的相似度 sim_{11}、“电脑”与“价格”的相似度 sim_{12}、“多少”与“计算机”的相似度 sim_{21}、“多少”与“价格”的相似度 sim_{22}、“钱”与“计算机”的相似度 sim_{31}、“钱”与“价格”的相似度 sim_{32}。取每个待匹配词与目标词集合的目标词计算得到的最大第二子相似度相乘，得到第二相似度 $sim2$。比如说，与“电脑”、“多少”、“钱”所对应的最大第二子相似度分别为 sim_{11}、sim_{22}、sim_{32}，则第二相似度 $sim2$ 可通过下列相乘公式计算得到 $sim2=sim_{11}\times sim_{22}\times sim_{32}$。

步骤 422，根据第一相似度和第二相似度进行计算，得到待匹配字符序列和目标字符序列的文本相似度。

举例来说，计算得到第一相似度 $sim1$ 和第二相似度 $sim2$ 之后，可将第一相似度 $sim1$ 和第二相似度 $sim2$ 乘以相应的第一权值 $w1$ 与第二权值 $w2$，计算得到文本相似度 $sim(S,T)=sim1\times w1+sim2\times w2$。

上述文本相似度计算方法中，在获取待匹配字符序列和目标字符序列之后，通过对匹配字符序列和目标字符序列进行预处理，得到以词语为单位按顺序形成的待匹配词序列和目标词序列，通过考虑词语顺序的第一相似度算法计算得到第一相似度，再根据待匹配词序列中包含的待匹配词和目标词序列中包含的目标词分别形成待匹配词集合与目标词集合，通过不考虑词语顺序的第二相似度算法计算得到第二相似度，然后综合第一相似度和第二相似度计算得到待匹配字符序列和目标字符序列之间的文本相似度。通过以词语为单位进行相似度计算，并综合两种相似度算法计算文本相似度，降低了以单字符通过单一相似度算法导致的误差，提高了文本相似度计算的准确性。

应该理解的是，虽然图 2 和 4 的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，这些步骤可以以其它的顺序执行。而且，图 2 和 4 中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，这些子步骤或者阶段的执行顺序也不必然是依次进行，而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

在其中一个实施例中，如图 5 所示，提供了一种文本相似度计算装置 500，该装置包括：字符序列获取模块 502，用于获取待匹配字符序列和目标字符序列；词序列获取模块 504，用于对待匹配字符序列和目标字符序列分别进行预处理，得到相应的待匹配词序列和目标词序列；第一相似度计算模块 506，用于将待匹配词序列中包含的待匹配词和目标词序列中包含的目标词通过第一相似度算法进行计算，得到第一相似度；词集合形成模块 508，用于提取所有待匹配词形成待匹配词集合，并提取所有目标词形成目标词集合；第二相似度计算模块 510，用于将待匹配词集合和目标词集合通过第二相似度算法进行计算，得到第二相似度；文本相似度计算模块 512，用于根据第一相似度和第二相似度进行计算，得到待匹配字符序列和目标字符序列的文本相似度。

在其中一个实施例中，词序列获取模块 504 还用于将待匹配字符序列中包含的无关字符和目标字符序列中包含的无关字符删除；对删除无关字符后的待匹配字符序列和目标字符序列分别进行分词，得到相应的待匹配词序列和目标词序列。

在其中一个实施例中，无关字符包括停用字符和相同字符；词序列获取模块 504 还用于将待匹配字符序列中包含的停用字符和目标字符序列中包含的停用字符删除；判断删除停用字符后的待匹配字符序列和目标字符序列中是否存在相同字符；相同字符是指在删除停用字符后的待匹配字符序列和目标字符序列中，处于相同位置的相同字符；若是，则将删除停用字符后的待匹配字符序列和目标字符序列中包含的相同字符删除，得到相应的待

匹配词序列和目标词序列。

在其中一个实施例中，第一相似度计算模块 506 还用于将待匹配词序列中包含的待匹配词和目标词序列中包含的目标词通过编辑距离公式进行计算，得到待匹配词序列与目标词序列之间的编辑距离；获取待匹配词序列中包含的待匹配词的第一数量，和目标词序列中包含的目标词的第二数量；根据编辑距离、第一数量和第二数量进行计算，得到第一相似度。

在其中一个实施例中，第二相似度计算模块 510 还用于将待匹配词集合和目标词集合进行匹配，统计待匹配词与目标词的匹配数量；统计待匹配词集合的待匹配词数量和目标词集合的目标词数量；根据匹配数量、待匹配词数量和目标词数量进行计算，得到第二相似度。

在其中一个实施例中，该装置还包括第三相似度计算模块 514，用于获取待匹配字符序列对应的待匹配拼音序列和目标字符序列对应的目标拼音序列；将待匹配拼音序列中包含的待匹配拼音和目标拼音序列中包含的目标拼音通过第一相似度算法进行计算，得到第三相似度；根据第一相似度和第二相似度进行计算，得到待匹配字符序列和目标字符序列的文本相似度，包括：根据第一相似度、第二相似度和第三相似度进行计算，得到待匹配字符序列和目标字符序列的文本相似度。

在其中一个实施例中，字符序列获取模块 502，还用于接收终端发送的待匹配字符序列；根据待匹配字符序列从数据库中获取多个目标字符序列；该装置还包括相关资源发送模块，用于查询文本相似度大于预设相似度阈值的目标字符序列所对应的相关资源；将相关资源发送至终端。

关于文本相似度计算装置的具体限定可以参见上文中对于文本相似度计算方法的限定，在此不再赘述。上述文本相似度计算装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，以便于处理器调用执行以上各个模块对应的操作。

在其中一个实施例中，提供了一种计算机设备，该计算机设备可以是服务器，其内部结构图可以如图 6 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。该计算机设备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计算机可读存储介质、内存储器。该非易失性计算机可读存储介质存储有操作系统、计算机可读指令和数据库。该内存储器为非易失性计算机可读存储介质中的操作系统和计算机可读指令的运行提供环境。该计算机设备的数据库用于存储目标字符序列。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。该计算机可读指令被处理器执行时以实现一种文本相似度计算方法。

本领域技术人员可以理解，图 6 中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可

以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

在其中一个实施例中，提供了一种计算机设备，包括存储器和一个或多个处理器，存储器中存储有计算机可读指令，计算机可读指令被处理器执行时实现本申请任意一个实施例中提供的文本相似度计算方法的步骤。

在其中一个实施例中，提供了一个或多个存储有计算机可读指令的非易失性计算机可读存储介质，计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器实现本申请任意一个实施例中提供的文本相似度计算方法的步骤。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机可读指令来指令相关的硬件来完成，所述的计算机可读指令可存储于一非易失性计算机可读取存储介质中，该计算机可读指令在执行时，可包括如上述各方法的实施例的流程。本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器（ROM）、可编程ROM（PROM）、电可编程ROM（EPROM）、电可擦除可编程ROM（EEPROM）或闪存。易失性存储器可包括随机存取存储器（RAM）或者外部高速缓冲存储器。作为说明而非局限，RAM以多种形式可得，诸如静态RAM（SRAM）、动态RAM（DRAM）、同步DRAM（SDRAM）、双数据率SDRAM（DDRSDRAM）、增强型SDRAM（ESDRAM）、同步链路（Synchlink）DRAM（SLDRAM）、存储器总线（Rambus）直接RAM（RDRAM）、直接存储器总线动态RAM（DRDRAM）、以及存储器总线动态RAM（RDRAM）等。

以上实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

权利要求书

1、一种文本相似度计算方法，包括:

获取待匹配字符序列和目标字符序列;

对所述待匹配字符序列和所述目标字符序列分别进行预处理，得到相应的待匹配词序列和目标词序列;

将所述待匹配词序列中包含的待匹配词和所述目标词序列中包含的目标词通过第一相似度算法进行计算，得到第一相似度;

提取所有待匹配词形成待匹配词集合，并提取所有目标词形成目标词集合;

将所述待匹配词集合和所述目标词集合通过第二相似度算法进行计算，得到第二相似度; 及

根据所述第一相似度和所述第二相似度进行计算，得到所述待匹配字符序列和所述目标字符序列的文本相似度。

2、根据权利要求 1 所述的方法，其特征在于，所述对所述待匹配字符序列和所述目标字符序列分别进行预处理，得到相应的待匹配词序列和目标词序列，包括:

将所述待匹配字符序列中包含的无关字符和所述目标字符序列中包含的无关字符删除; 及

对删除无关字符后的待匹配字符序列和目标字符序列分别进行分词，得到相应的待匹配词序列和目标词序列。

3、根据权利要求 2 所述的方法，其特征在于，所述无关字符包括停用字符和相同字符; 所述将所述待匹配字符序列中包含的无关字符和所述目标字符序列中包含的无关字符删除，包括:

将所述待匹配字符序列中包含的停用字符和所述目标字符序列中包含的停用字符删除; 及

判断删除停用字符后的待匹配字符序列和目标字符序列中是否存在相同字符; 所述相同字符是指在所述删除停用字符后的待匹配字符序列和目标字符序列中，处于相同位置的相同字符;

若是，则将所述删除停用字符后的待匹配字符序列和目标字符序列中包含的所述相同字符删除，得到相应的待匹配词序列和目标词序列。

4、根据权利要求 1 所述的方法，其特征在于，所述将所述待匹配词序列中包含的待匹配词和所述目标词序列中包含的目标词通过第一相似度算法进行计算，得到第一相似度，包括:

将所述待匹配词序列中包含的待匹配词和所述目标词序列中包含的目标词通过编辑距离公式进行计算，得到所述待匹配词序列与所述目标词序列之间的编辑距离;

获取所述待匹配词序列中包含的待匹配词的第一数量，和所述目标词序列中包含的目标词的第二数量; 及

根据所述编辑距离、第一数量和第二数量进行计算，得到第一相似度。

5、根据权利要求 1 所述的方法，其特征在于，所述将所述待匹配词集合和所述目标词集合通过第二相似度算法进行计算，得到第二相似度，包括:

将所述待匹配词集合和所述目标词集合进行匹配，统计所述待匹配词与所述目标词的匹配数量;

统计所述待匹配词集合的待匹配词数量和所述目标词集合的目标词数量；及

根据所述匹配数量、待匹配词数量和目标词数量进行计算，得到第二相似度。

6、根据权利要求 1 至 5 任意一项所述的方法，其特征在于，在所述获取待匹配字符序列和目标字符序列之后，还包括:

获取所述待匹配字符序列对应的待匹配拼音序列和目标字符序列对应的目标拼音序列;

将所述待匹配拼音序列中包含的待匹配拼音和所述目标拼音序列中包含的目标拼音通过第一相似度算法进行计算，得到第三相似度;

所述根据所述第一相似度和所述第二相似度进行计算，得到所述待匹配字符序列和所述目标字符序列的文本相似度，包括：及

根据所述第一相似度、所述第二相似度和第三相似度进行计算，得到所述待匹配字符序列和所述目标字符序列的文本相似度。

7、根据权利要求 1 至 5 任意一项所述的方法，其特征在于，所述获取待匹配字符序列和目标字符序列，包括:

接收终端发送的待匹配字符序列;

根据所述待匹配字符序列从数据库中获取多个目标字符序列;

在所述根据所述第一相似度和所述第二相似度进行计算，得到所述待匹配字符序列和所述目标字符序列的文本相似度之后，还包括:

查询文本相似度大于预设相似度阈值的目标字符序列所对应的相关资源；及

将所述相关资源发送至所述终端。

8、一种文本相似度计算装置，包括:

字符序列获取模块，用于获取待匹配字符序列和目标字符序列;

词序列获取模块，用于对所述待匹配字符序列和所述目标字符序列分别进行预处理，得到相应的待匹配词序列和目标词序列;

第一相似度计算模块，用于将所述待匹配词序列中包含的待匹配词和所述目标词序列中包含的目标词通过第一相似度算法进行计算，得到第一相似度;

词集合形成模块，用于提取所有待匹配词形成待匹配词集合，并提取所有目标词形成目标词集合;

第二相似度计算模块，用于将所述待匹配词集合和所述目标词集合通过第二相似度算法进行计算，得到第二相似度；及

文本相似度计算模块，用于根据所述第一相似度和所述第二相似度进行计算，得到所述待匹配字符序列和所述目标字符序列的文本相似度。

9、一种计算机设备，包括存储器及一个或多个处理器，所述存储器中储存有计算机可读指令，所述计算机可读指令被所述一个或多个处理器执行时，使得所述一个或多个处理器执行以下步骤：

获取待匹配字符序列和目标字符序列；

对所述待匹配字符序列和所述目标字符序列分别进行预处理，得到相应的待匹配词序列和目标词序列；

将所述待匹配词序列中包含的待匹配词和所述目标词序列中包含的目标词通过第一相似度算法进行计算，得到第一相似度；

提取所有待匹配词形成待匹配词集合，并提取所有目标词形成目标词集合；

将所述待匹配词集合和所述目标词集合通过第二相似度算法进行计算，得到第二相似度；及

根据所述第一相似度和所述第二相似度进行计算，得到所述待匹配字符序列和所述目标字符序列的文本相似度。

10、根据权利要求 9 所述的计算机设备，其特征在于，所述对所述待匹配字符序列和所述目标字符序列分别进行预处理，得到相应的待匹配词序列和目标词序列的步骤，包括执行以下步骤：

将所述待匹配字符序列中包含的无关字符和所述目标字符序列中包含的无关字符删除；及

对删除无关字符后的待匹配字符序列和目标字符序列分别进行分词，得到相应的待匹配词序列和目标词序列。

11、根据权利要求 10 所述的计算机设备，其特征在于，所述无关字符包括停用字符和相同字符；所述将所述待匹配字符序列中包含的无关字符和所述目标字符序列中包含的无关字符删除的步骤，包括执行以下步骤：

将所述待匹配字符序列中包含的停用字符和所述目标字符序列中包含的停用字符删除；及

判断删除停用字符后的待匹配字符序列和目标字符序列中是否存在相同字符；所述相同字符是指在所述删除停用字符后的待匹配字符序列和目标字符序列中，处于相同位置的相同字符；

若是，则将所述删除停用字符后的待匹配字符序列和目标字符序列中包含的所述相同字符删除，得到相应的待匹配词序列和目标词序列。

12、根据权利要求 9 所述的计算机设备，其特征在于，所述将所述待匹配词序列中包含的待匹配词和所述目标词序列中包含的目标词通过第一相似度算法进行计算，得到第一相似度的步骤，包括执行以下步骤：

将所述待匹配词序列中包含的待匹配词和所述目标词序列中包含的目标词通过编辑距离公式进行计算，得到所述待匹配词序列与所述目标词序列之间的编辑距离；

获取所述待匹配词序列中包含的待匹配词的第一数量，和所述目标词序列中包含的目标词的第二数量；及

根据所述编辑距离、第一数量和第二数量进行计算，得到第一相似度。

13 根据权利要求 9 所述的计算机设备，其特征在于，所述将所述待匹配词集合和所述目标词集合通过第二相似度算法进行计算，得到第二相似度的步骤，包括执行以下步骤：

将所述待匹配词集合和所述目标词集合进行匹配，统计所述待匹配词与所述目标词的匹配数量；

统计所述待匹配词集合的待匹配词数量和所述目标词集合的目标词数量；及

根据所述匹配数量、待匹配词数量和目标词数量进行计算，得到第二相似度。

14、根据权利要求 9 至 13 任意一项所述的计算机设备，其特征在于，在所述获取待匹配字符序列和目标字符序列的步骤之后，还包括执行以下步骤：

获取所述待匹配字符序列对应的待匹配拼音序列和目标字符序列对应的目标拼音序列；

将所述待匹配拼音序列中包含的待匹配拼音和所述目标拼音序列中包含的目标拼音通过第一相似度算法进行计算，得到第三相似度；

所述根据所述第一相似度和所述第二相似度进行计算，得到所述待匹配字符序列和所述目标字符序列的文本相似度，包括：及

根据所述第一相似度、所述第二相似度和第三相似度进行计算，得到所述待匹配字符序列和所述目标字符序列的文本相似度。

15、一个或多个存储有计算机可读指令的非易失性计算机可读存储介质，所述计算机可读指令被一个或多个处理器执行时，使得所述一个或多个处理器执行以下步骤：

获取待匹配字符序列和目标字符序列；

对所述待匹配字符序列和所述目标字符序列分别进行预处理，得到相应的待匹配词序列和目标词序列；

将所述待匹配词序列中包含的待匹配词和所述目标词序列中包含的目标词通过第一相似度算法进行计算，得到第一相似度；

提取所有待匹配词形成待匹配词集合，并提取所有目标词形成目标词集合；

将所述待匹配词集合和所述目标词集合通过第二相似度算法进行计算，得到第二相似度；及

根据所述第一相似度和所述第二相似度进行计算，得到所述待匹配字符序列和所述目标字符序列的文本相似度。

16、根据权利要求 15 所述的存储介质，其特征在于，所述对所述待匹配字符序列和所述目标字符序列分别进行预处理，得到相应的待匹配词序列和目标词序列的步骤，包括

执行以下步骤:

将所述待匹配字符序列中包含的无关字符和所述目标字符序列中包含的无关字符删除; 及

对删除无关字符后的待匹配字符序列和目标字符序列分别进行分词, 得到相应的待匹配词序列和目标词序列。

17、根据权利要求 16 所述的存储介质, 其特征在于, 所述无关字符包括停用字符和相同字符; 所述将所述待匹配字符序列中包含的无关字符和所述目标字符序列中包含的无关字符删除的步骤, 包括执行以下步骤:

将所述待匹配字符序列中包含的停用字符和所述目标字符序列中包含的停用字符删除; 及

判断删除停用字符后的待匹配字符序列和目标字符序列中是否存在相同字符; 所述相同字符是指在所述删除停用字符后的待匹配字符序列和目标字符序列中, 处于相同位置的相同字符;

若是, 则将所述删除停用字符后的待匹配字符序列和目标字符序列中包含的所述相同字符删除, 得到相应的待匹配词序列和目标词序列。

18、根据权利要求 15 所述的存储介质, 其特征在于, 所述将所述待匹配词序列中包含的待匹配词和所述目标词序列中包含的目标词通过第一相似度算法进行计算, 得到第一相似度的步骤, 包括执行以下步骤:

将所述待匹配词序列中包含的待匹配词和所述目标词序列中包含的目标词通过编辑距离公式进行计算, 得到所述待匹配词序列与所述目标词序列之间的编辑距离;

获取所述待匹配词序列中包含的待匹配词的第一数量, 和所述目标词序列中包含的目标词的第二数量; 及

根据所述编辑距离、第一数量和第二数量进行计算, 得到第一相似度。

19 根据权利要求 15 所述的存储介质, 其特征在于, 所述将所述待匹配词集合和所述目标词集合通过第二相似度算法进行计算, 得到第二相似度的步骤, 包括执行以下步骤:

将所述待匹配词集合和所述目标词集合进行匹配, 统计所述待匹配词与所述目标词的匹配数量;

统计所述待匹配词集合的待匹配词数量和所述目标词集合的目标词数量; 及

根据所述匹配数量、待匹配词数量和目标词数量进行计算, 得到第二相似度。

20、根据权利要求 15 至 19 任意一项所述的存储介质, 其特征在于, 在所述获取待匹配字符序列和目标字符序列的步骤之后, 还包括执行以下步骤:

获取所述待匹配字符序列对应的待匹配拼音序列和目标字符序列对应的目标拼音序列;

将所述待匹配拼音序列中包含的待匹配拼音和所述目标拼音序列中包含的目标拼音通过第一相似度算法进行计算, 得到第三相似度;

所述根据所述第一相似度和所述第二相似度进行计算，得到所述待匹配字符序列和所述目标字符序列的文本相似度，包括：及

根据所述第一相似度、所述第二相似度和第三相似度进行计算，得到所述待匹配字符序列和所述目标字符序列的文本相似度。

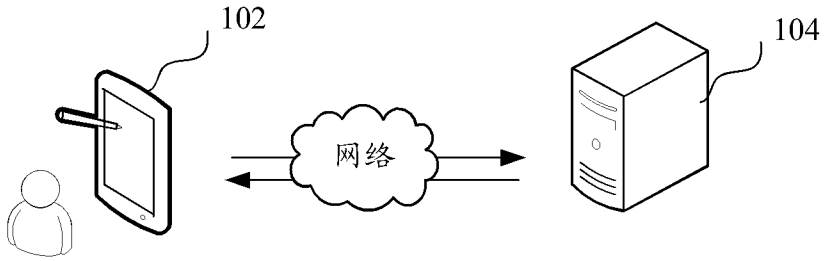

102
104
网络

图 1


202
获取待匹配字符序列和目标字符序列
204
对待匹配字符序列和目标字符序列分别进行预处理，得到相应的待匹配词序列和目标词序列
206
将待匹配词序列中包含的待匹配词和目标词序列中包含的目标词通过第一相似度算法进行计算，得到第一相似度
208
提取所有待匹配词形成待匹配词集合，并提取所有目标词形成目标词集合
210
将待匹配词集合和目标词集合通过第二相似度算法进行计算，得到第二相似度
212
根据第一相似度和第二相似度进行计算，得到待匹配字符序列和目标字符序列的文本相似度

图 2

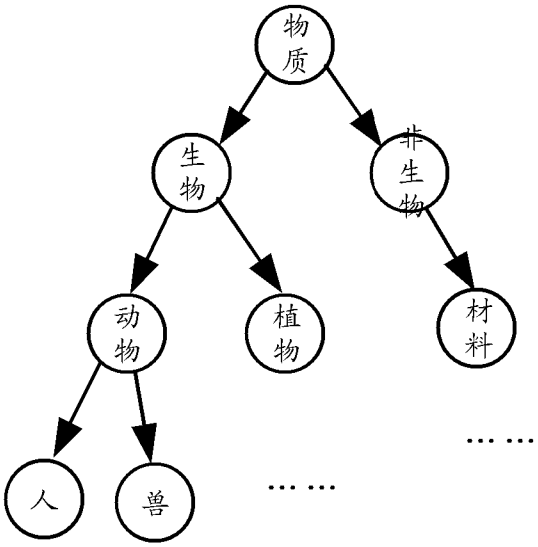


图 3A

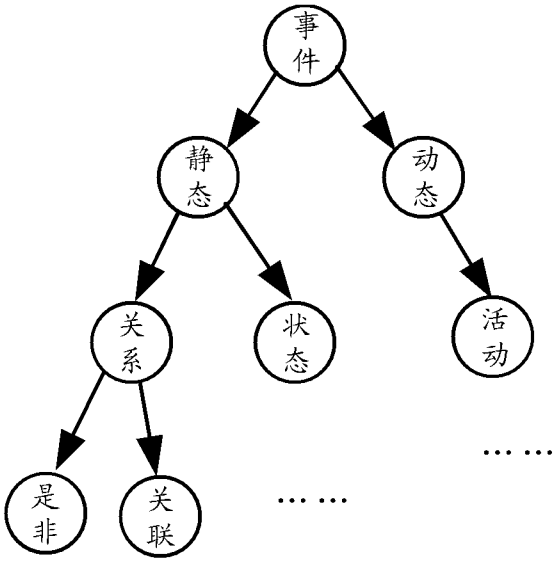


图 3B

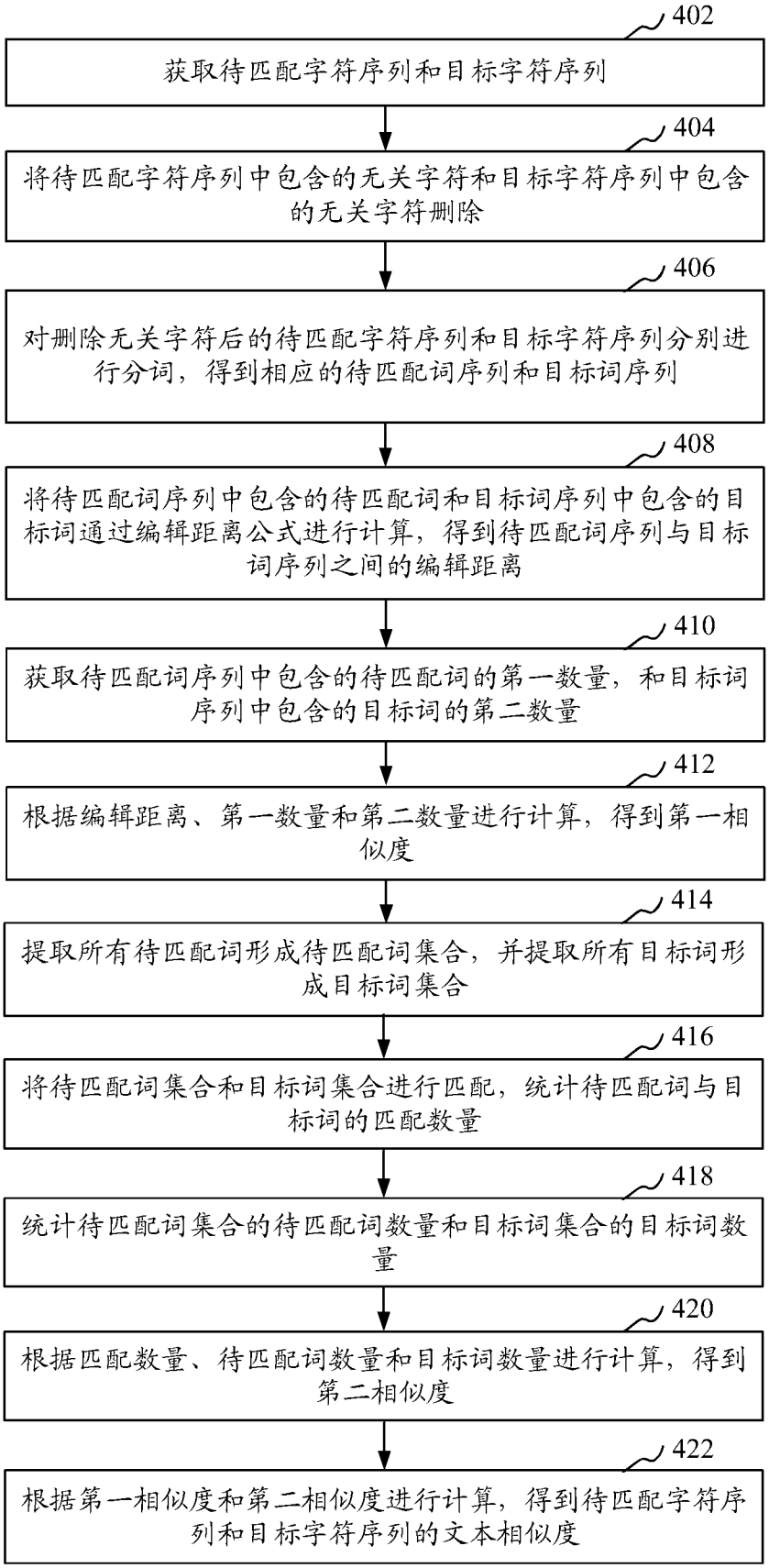

402
获取待匹配字符序列和目标字符序列
404
将待匹配字符序列中包含的无关字符和目标字符序列中包含的无关字符删除
406
对删除无关字符后的待匹配字符序列和目标字符序列分别进行分词，得到相应的待匹配词序列和目标词序列
408
将待匹配词序列中包含的待匹配词和目标词序列中包含的目标词通过编辑距离公式进行计算，得到待匹配词序列与目标词序列之间的编辑距离
410
获取待匹配词序列中包含的待匹配词的第一数量，和目标词序列中包含的目标词的第二数量
412
根据编辑距离、第一数量和第二数量进行计算，得到第一相似度
414
提取所有待匹配词形成待匹配词集合，并提取所有目标词形成目标词集合
416
将待匹配词集合和目标词集合进行匹配，统计待匹配词与目标词的匹配数量
418
统计待匹配词集合的待匹配词数量和目标词集合的目标词数量
420
根据匹配数量、待匹配词数量和目标词数量进行计算，得到第二相似度
422
根据第一相似度和第二相似度进行计算，得到待匹配字符序列和目标字符序列的文本相似度

图 4

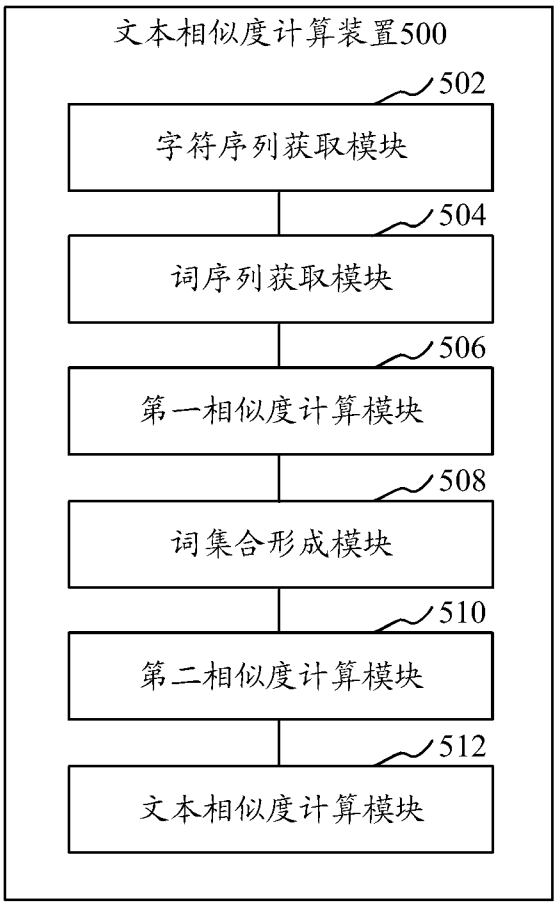

文本相似度计算装置500
502
字符序列获取模块
504
词序列获取模块
506
第一相似度计算模块
508
词集合形成模块
510
第二相似度计算模块
512
文本相似度计算模块

图 5

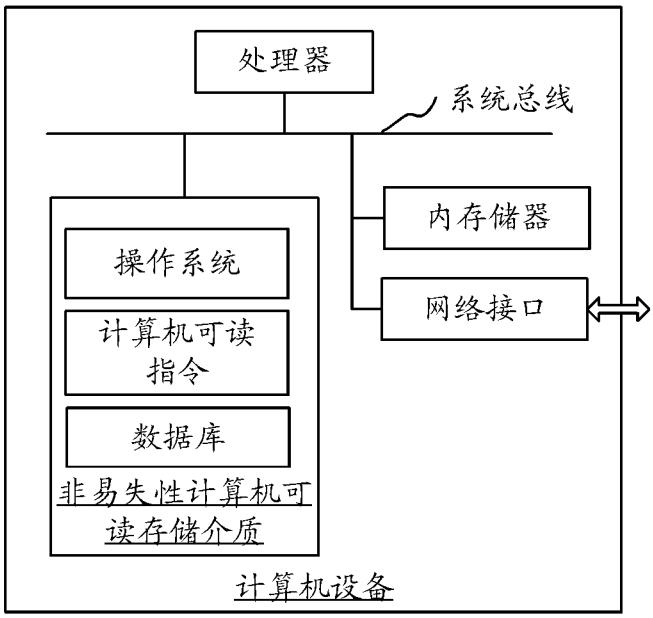

处理器
系统总线
内存储器
网络接口
操作系统
计算机可读指令
数据库
非易失性计算机可读存储介质
计算机设备

图 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/099994

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/27(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)

CNABS; CNTXT; TWTXT; TWABS; DWPI; VEN; USTXT; EPTXT; WOTXT; CNKI: 文本, 相似度, 匹配, 字符, 词序, 顺序, 目标, text, similarity, match, character, word order, sequence, goal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 108304378 A (ONE CONNECT SMART TECHNOLOGY CO., LTD. (SHENZHEN)) 20 July 2018 (2018-07-20) entire document	1-20
X	CN 107491425 A (HEFEI MIDEA INTELLIGENT TECHNOLOGY CO., LTD.) 19 December 2017 (2017-12-19) description, paragraphs 7-43	1-20
A	CN 103176962 A (SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY) 26 June 2013 (2013-06-26) entire document	1-20
A	US 2009175545 A1 (XEROX CORP.) 09 July 2009 (2009-07-09) 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

"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	Date of mailing of the international search report
19 October 2018	**01 November 2018**

Name and mailing address of the ISA/CN	Authorized officer
State Intellectual Property Office of the P. R. China **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing** **100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/099994

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	108304378	A	20 July 2018	None			
CN	107491425	A	19 December 2017	None			
CN	103176962	A	26 June 2013	CN	103176962	B	04 November 2015
US	2009175545	A1	09 July 2009	US	8077984	B2	13 December 2011

国际检索报告

国际申请号 PCT/CN2018/099994

A. 主题的分类

G06F 17/27(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）)

CNABS;CNTXT;TWTXT;TWABS;DWPI;VEN;USTXT;EPTXT;WOTXT;CNKI:文本，相似度，匹配，字符，词序，顺序，目标，text, similarity, match, character, word order, sequence, goal

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
PX	CN 108304378 A (深圳壹账通智能科技有限公司) 2018年 7月 20日 (2018-07-20) 全文	1-20
X	CN 107491425 A (合肥美的智能科技有限公司) 2017年 12月 19日 (2017-12-19) 说明书第7-43段	1-20
A	CN 103176962 A (深圳先进技术研究院) 2013年 6月 26日 (2013-06-26) 全文	1-20
A	US 2009175545 A1 (XEROX CORP) 2009年 7月 9日 (2009-07-09) 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	国际检索报告邮寄日期
2018年 10月 19日	2018年 11月 1日

ISA/CN的名称和邮寄地址	受权官员
中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088	房琦
传真号 (86-10)62019451	电话号码 86-(20)-28950714

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/099994

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	108304378	A	2018年 7月 20日	无			
CN	107491425	A	2017年 12月 19日	无			
CN	103176962	A	2013年 6月 26日	CN	103176962	B	2015年 11月 4日
US	2009175545	A1	2009年 7月 9日	US	8077984	B2	2011年 12月 13日