

# (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 11 月 28 日 (28.11.2019)



(10) 国际公布号  
**WO 2019/223103 A1**

(51) 国际专利分类号:  
**G06F 17/27** (2006.01)

(21) 国际申请号: PCT/CN2018/097497

(22) 国际申请日: 2018 年 7 月 27 日 (27.07.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:  
201810495130.5 2018年5月22日 (22.05.2018) 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: 李育儒(LI, Yuru); 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)。王鸿滨(WANG, Hongbin); 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)。吴晓贝(WU, Xiaobei); 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)。汪伟(WANG, Wei); 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: 深圳中一专利商标事务所(SHENZHEN ZHONGYI PATENT AND TRADEMARK OFFICE); 中国广东省深圳市福田区深南中路

(54) Title: TEXT SIMILARITY ACQUISITION METHOD AND APPARATUS, TERMINAL DEVICE AND MEDIUM

(54) 发明名称: 文本相似度的获取方法、装置、终端设备及介质

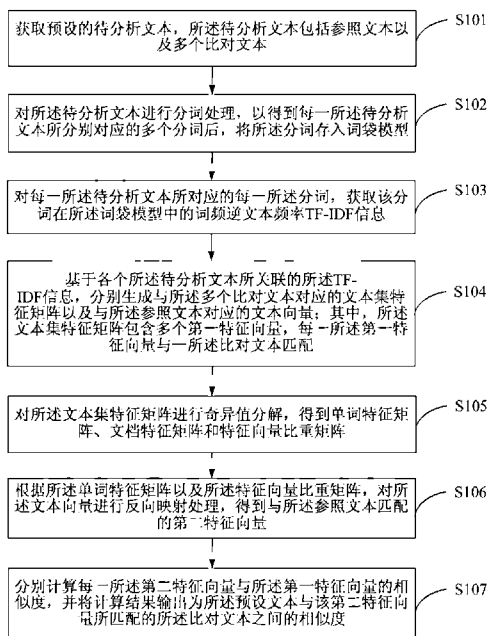


图 1

S101 ACQUIRE PRESET TEXT TO BE ANALYZED, THE TEXT TO BE ANALYZED COMPRISING REFERENCE TEXT AND A PLURALITY OF PIECES OF COMPARISON TEXT

S102 PERFORM WORD SEGMENTATION PROCESSING ON THE TEXT TO BE ANALYZED TO OBTAIN A PLURALITY OF SEGMENTED WORDS CORRESPONDING TO EACH PIECE OF THE TEXT TO BE ANALYZED, AND STORE THE SEGMENTED WORDS INTO A BAG-OF-WORDS MODEL

S103 FOR EACH OF SEGMENTED WORDS CORRESPONDING TO EACH PIECE OF THE TEXT TO BE ANALYZED, ACQUIRE TERM FREQUENCY INVERSE DOCUMENT FREQUENCY (TF-IDF) INFORMATION OF EACH OF THE SEGMENTED WORDS IN THE BAG-OF-WORDS MODEL

S104 ON THE BASIS OF TF-IDF INFORMATION ASSOCIATED WITH THE TEXT TO BE ANALYZED, RESPECTIVELY GENERATE A TEXT SET FEATURE MATRIX CORRESPONDING TO THE PLURALITY OF PIECES OF COMPARISON TEXT AND TEXT VECTORS CORRESPONDING TO THE REFERENCE TEXT; THE TEXT SET FEATURE MATRIX COMPRISING A PLURALITY OF FIRST FEATURE VECTORS, EACH OF THE FIRST FEATURE VECTORS MATCHING THE COMPARISON TEXT

S105 PERFORM SINGULAR VALUE DECOMPOSITION ON THE TEXT SET FEATURE MATRIX, TO OBTAIN A WORD FEATURE MATRIX, A DOCUMENT FEATURE MATRIX AND A FEATURE VECTOR PROPORTION MATRIX

S106 PERFORM REVERSE MAPPING PROCESSING ON THE TEXT VECTORS ACCORDING TO THE WORD FEATURE MATRIX AND THE FEATURE VECTOR PROPORTION MATRIX, TO OBTAIN SECOND FEATURE VECTORS MATCHING THE REFERENCE TEXT

S107 CALCULATE THE SIMILARITY BETWEEN EACH OF THE SECOND FEATURE VECTORS AND THE FIRST FEATURE VECTOR RESPECTIVELY, AND OUTPUT CALCULATION RESULTS AS THE SIMILARITIES BETWEEN THE PRESET TEXT AND THE COMPARISON TEXT MATCHING THE SECOND FEATURE VECTORS

(57) Abstract: The present solution provides a text similarity acquisition method and apparatus, a terminal device and a medium, applicable to the field of data processing technology. The method comprises: upon acquisition of a plurality of segmented words respectively corresponding to each piece of text to be analyzed, storing the segmented words into a bag-of-words model; acquiring TF-IDF information of each of the segmented words in the bag-of-words model; on the basis of TF-IDF information associated with the text to be analyzed, respectively generating a text set feature matrix corresponding to a plurality of pieces of comparison text and text vectors corresponding to reference text; performing singular value decomposition on the text set feature matrix, and performing reverse

1014 号老特区报社四楼 (5 号信箱),  
Guangdong 518028 (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, 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))。

mapping processing on the text vectors according to an obtained word feature matrix and a feature vector proportion matrix to obtain second feature vectors; calculating the similarity between each of the second feature vectors and a first feature vector respectively, and outputting calculation results as the similarities between the preset text and the comparison text matching the second feature vectors. This solution improves the accuracy of calculation of text similarity and the efficiency of text comparison.

(57) 摘要: 本方案提供了一种文本相似度的获取方法、装置、终端设备及介质, 适用于数据处理技术领域, 该方法包括: 获取每一待分析文本所分别对应的多个分词后, 将分词存入词袋模型; 获取每一分词在词袋模型中的TF-IDF信息; 基于各个待分析文本所关联的TF-IDF信息, 分别生成与多个比对文本对应的文本集特征矩阵以及与参照文本对应的文本向量; 对文本集特征矩阵进行奇异值分解, 并根据得到的单词特征矩阵以及特征向量比重矩阵, 对文本向量进行反向映射处理, 得到第二特征向量; 分别计算每一第二特征向量与第一特征向量的相似度, 并将计算结果输出为预设文本与该第二特征向量所匹配的比对文本之间的相似度。本方案提高了文本相似度的计算准确率以及文本比对效率。

## 文本相似度的获取方法、装置、终端设备及介质

本申请要求于 2018 年 05 月 22 日提交中国专利局、申请号为 201810495130.5、发明名称为“文本相似度的获取方法、终端设备及介质”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。

### 5 技术领域

本申请属于数据处理技术领域，尤其涉及一种文本相似度的获取方法、装置、终端设备及计算机可读存储介质。

### 背景技术

文本相似度用于度量文本与文本之间的相似程度。传统方式中，可通过人工判断的方式来确定文本相似度。然而，人工判断大量相似文本常常是一个耗时且繁琐的工作，因此，为了解决这一问题，随着科学研究的不断发展，衍生出了词频统计以及 simhash 等向量空间模型。这些向量空间模型通过识别两篇文章所共同存在的单词，并基于单词的出现与否及各单词的词频等信息来计算文本相似度。因此，仅当两篇文章均存在数目较多且完全相同的单词时，计算得到的文本相似度才会相对较高。

然而，单词具有多义性和同义性，即，用于表示相同语义的单词在表达形式上可能不同，而相同表达形式的单词也可能表达出不同的语义，故基于上述向量空间模型来计算文本相似度时，可能会出现两篇文章实际表达的意思相同（文本相似度理应非常高），但计算得到的文本相似度却较为低下的情况。因此，现有的文本相似度计算方法存在计算准确率较低的问题。

### 20 技术问题

有鉴于此，本申请实施例提供了一种文本相似度的获取方法、装置、终端设备及计算机可读存储介质，以解决现有技术中文本相似度的计算准确率较为低下的问题。

### 技术解决方案

本申请实施例的第一方面提供了一种文本相似度的获取方法，包括：

25 获取预设的待分析文本，所述待分析文本包括参照文本以及多个比对文本；

对所述待分析文本进行分词处理，以得到每一所述待分析文本所分别对应的多个分词后，将所述分词存入词袋模型；

对每一所述待分析文本所对应的每一所述分词，获取该分词在所述词袋模型中的词频逆文本频率 TF-IDF 信息；

30 基于各个所述待分析文本所关联的所述 TF-IDF 信息，分别生成与所述多个比对文本对

应的文本集特征矩阵以及与所述参照文本对应的文本向量；其中，所述文本集特征矩阵包含多个第一特征向量，每一所述第一特征向量与一所述比对文本匹配；

对所述文本集特征矩阵进行奇异值分解，得到单词特征矩阵、文档特征矩阵和特征向量比重矩阵；

5 根据所述单词特征矩阵以及所述特征向量比重矩阵，对所述文本向量进行反向映射处理，得到与所述参照文本匹配的第二特征向量；

分别计算每一所述第二特征向量与所述第一特征向量的相似度，并将计算结果输出为所述参照文本与该第二特征向量所匹配的所述比对文本之间的相似度。

本申请实施例的第二方面提供了一种文本相似度的获取装置，所述识别装置包括用于执行上述第一方面所述的文本相似度的获取方法的单元。

本申请实施例的第三方面提供了一种终端设备，包括存储器以及处理器，所述存储器中存储有可在所述处理器上运行的计算机可读指令，所述处理器执行所述计算机可读指令时实现如第一方面所述的文本相似度的获取方法的步骤。

本申请实施例的第四方面提供了一种计算机可读存储介质，所述计算机可读存储介质存储有计算机可读指令，所述计算机可读指令被处理器执行时实现如第一方面所述的文本相似度的获取方法的步骤。

有益效果

本申请实施例中，通过生成参照文本所关联的文本向量以及生成多个比对文本所共同关联的文本集特征矩阵，基于对文本集特征矩阵的奇异值分解，并计算分解后用于表征各待分析文本的特征向量之间的相似度，保证了计算得出的文本相似度能够较为准确地体现文本之间的主题相似性，而并非仅仅依赖于文本之间共同存在的单词来计算相似度，因此，提高了文本相似度的计算准确率。并且，在一次运算过程中，本申请实施例能够同时计算参照文本分别与多个对比文本之间的相似度，故也提高了文本的比对效率。

附图说明

25 图 1 是本申请实施例提供的文本相似度的获取方法的实现流程图；

图 2 是本申请实施例提供的基于多个比对文本所生成的文本集特征矩阵的示意图；

图 3 是本申请另一实施例提供的文本相似度的获取方法的实现流程图；

图 4 是本申请另一实施例提供的文本相似度的获取方法 S109 的具体实现流程图；

图 5 是本申请另一实施例提供的文本相似度的获取方法 S1091 的具体实现流程图；

30 图 6 是本申请实施例提供的文本相似度的获取方法 S102 的具体实现流程图；

图 7 是本申请实施例提供的文本相似度的获取装置的结构框图；

图 8 是本申请实施例提供的终端设备的示意图。

## 本发明的实施方式

为了说明本申请所述的技术方案，下面通过具体实施例来进行说明。

图 1 示出了本申请实施例提供的信息的输入方法的实现流程，该方法流程包括步骤

5 S101 至 S107。各步骤的具体实现原理如下：

S101：获取预设的待分析文本，所述待分析文本包括参照文本以及多个比对文本。

本申请实施例中，参照文本为预先选定的一个文本。对用户而言，若其需要获取任一  
10 文本与参照文本之间的文本相似度，则所述的任一文本为比对文本。比对文本可以为一个，  
也可以为一个以上。优选地，将参照文本以及比对文本统称为待分析文本，且获取预设的一  
个参照文本以及一个以上的比对文本。

本申请实施例中，待分析文本包括但不限于从各类新闻网站中所收集得到的数据文本  
以及企业自身所发布的公开文本等。上述公开文本例如可以是财务年报、财务季报以及招股  
书。

15 S102：对所述待分析文本进行分词处理，以得到每一所述待分析文本所分别对应的的多  
个分词后，将所述分词存入词袋模型。

对获取到的每一待分析文本，通过预设的分词算法对该待分析文本进行分词处理，  
得到该待分析文本所对应的各个分词后，从中筛选出互不重复的分词，并将这些互不重复的  
分词输出至词袋模型。因此，对于当前时刻所获得的任一待分析文本，词袋模型中将包含有  
存在于该待分析文本中的各个分词。

20 S103：对每一所述待分析文本所对应的的每一所述分词，获取该分词在所述词袋模型中  
的词频逆文本频率 TF-IDF 信息。

S104：基于各个所述待分析文本所关联的所述 TF-IDF 信息，分别生成与所述多个比对  
文本对应的文本集特征矩阵以及与所述参照文本对应的文本向量；其中，所述文本集特征矩  
阵包含多个第一特征向量，每一所述第一特征向量与一所述比对文本匹配。

25 对于词袋模型中的每一个分词，确定该分词在每一待分析文本中所出现的频率，则上  
述频率即为该分词在词袋模型中的词频逆文本频率（TF-IDF，term frequency Inverse  
Document Frequency）。

本申请实施例中，生成每一比对文本对应的一个第一特征向量。具体地，将词袋模型  
中每一分词在比对文本中的 TF-IDF 信息输出为该比对文本所对应的第一特征向量中的一个  
30 元素。因此，第一特征向量所包含的元素数量与词袋模型所存储的分词数量相同。同理，生  
成与参照文本对应的一个文本向量，使得词袋模型中每一分词在参照文本中的 TF-IDF 信息

能够输出为文本向量中的一个元素。

示例性地，假设有三个比对文本，分别为 d1、d2 和 d3，参照文本为 q，且各待分析文本所包含的单词如下：

d1: Shipment of gold damaged in fire.

5 d2: Delivery of silver arrived in a silver truck.

d3: Shipment of gold arrived in a truck.

q: Gold silver truck.

则对上述各个待分析文本进行分词处理后，得到的不包含重复分词的词袋模型为：{a arrived damaged delivery fire gold in of shipment silver truck}。以 d1 为例，词袋模型所包含的  
10 各个分词在 d1 中的 TF-IDF 信息分别为 1、0、1、0、1、1、1、1、1、0 以及 0，因此，d1 所对应生成的第一特征向量为[1 0 1 0 1 1 1 1 1 0 0]；词袋模型所包含的各个分词在 q 中的 TF-IDF 信息分别为 0、0、0、0、0、1、0、0、0、1 以及 1，因此，q 所对应生成的文本向量为[0 0 0 0 0 1 0 0 0 1 1]。

本申请实施例中，将包含各个比对文本所对应的第一特征向量的集合作为上述文本集  
15 特征矩阵。其中，每一第一特征向量表示文本集特征矩阵中的一列矩阵元素。因此，上述示例中，与比对文本 d1、d2 以及 d3 所匹配的文本集特征矩阵 A 可参见图 2。

S105：对所述文本集特征矩阵进行奇异值分解，得到单词特征矩阵、文档特征矩阵和特征向量比重矩阵。

本申请实施例中，通过预设的奇异值分解算法，对文本集特征矩阵 A 进行奇异值分解  
20 处理，以求解得到单词特征矩阵 U、文档特征矩阵 V 和特征向量比重矩阵 S。

示例性地，基于  $A=USV^T$ ，可将文本集特征矩阵 A 拆分为多个矩阵正交相乘的形式，则对于上述图 2 所示的文本集特征矩阵 A，其分解得到的单词特征矩阵 U、文档特征矩阵 V 和特征向量比重矩阵 S 具体如下：

$$U = \begin{bmatrix} -0.4201 & 0.0748 & -0.0460 \\ -0.2995 & -0.2001 & 0.4078 \\ -0.1206 & 0.2749 & -0.4538 \\ -0.1576 & -0.3046 & -0.2006 \\ -0.1206 & 0.2749 & -0.4538 \\ -0.2626 & 0.3794 & 0.1547 \\ -0.4201 & 0.0748 & -0.0460 \\ -0.4201 & 0.0748 & -0.0460 \\ -0.2626 & 0.3794 & 0.1547 \\ -0.3151 & -0.6093 & -0.4013 \\ -0.2995 & -0.2001 & 0.4078 \end{bmatrix} \quad S = \begin{bmatrix} 4.0989 & 0.0000 & 0.0000 \\ 0.0000 & 2.3616 & 0.0000 \\ 0.0000 & 0.0000 & 1.2737 \end{bmatrix}$$

$$V = \begin{bmatrix} -0.4945 & 0.6492 & -0.5780 \\ -0.6458 & -0.7194 & -0.2556 \\ -0.5817 & 0.2469 & 0.7750 \end{bmatrix} V^T = \begin{bmatrix} -0.4945 & 0.6492 & -0.5780 \\ -0.6458 & -0.7194 & -0.2556 \\ -0.5817 & 0.2469 & 0.7750 \end{bmatrix}$$

S106: 根据所述单词特征矩阵以及所述特征向量比重矩阵, 对所述文本向量进行反向映射处理, 得到与所述参照文本匹配的第二特征向量。

本申请实施例中, 为了比较参照文本与各个比对文本的文本相似度, 在生成参照文本所对应的文本向量 $q$ 后, 将其映射到与文本集特征矩阵奇异值分解后相同的文档特征空间中。即, 利用相应的单词特征矩阵 $U$ 和特征向量比重矩阵 $S$ , 反向映射回文档特征空间 $V_q$ , 使得 $V_q = q^T U S^{-1}$ , 从而得到与参照文本匹配的第二特征向量。

S107: 分别计算每一所述第二特征向量与所述第一特征向量的相似度, 并将计算结果输出为所述参照文本与该第二特征向量所匹配的所述比对文本之间的相似度。

本申请实施例中, 读取文本集特征矩阵中的各个第一特征向量, 并基于读取得到的每一第一特征向量与参照文本所匹配的第二特征向量的相似度, 确定各个比对文本分别与参照文本之间的文本相似度。

示例性地, 第一特征向量与第二特征向量之间的相似度具体为余弦相似度。余弦相似度 $S$ 的计算公式具体如下:

$$S = \frac{\sum_{i=1}^n (x_i y_i)}{\sqrt{\sum_{i=1}^n (x_i)^2} \sqrt{\sum_{i=1}^n (y_i)^2}}$$

其中,  $x_i$ 表示第一特征向量的第 $i$ 个元素值;  $y_i$ 表示第二特征向量的第 $i$ 个元素值;  $n$ 表示第一特征向量或第二特征向量的元素总数。

由于每一个第一特征向量与一个比对文本对应, 故在计算任一第一特征向量与第二特征向量的余弦相似度后, 将该余弦相似度输出为该第一特征向量所对应的比对文本与参照文本之间的文本相似度。

本申请实施例中, 通过生成参照文本所关联的文本向量以及生成多个比对文本所共同关联的文本集特征矩阵, 基于对文本集特征矩阵的奇异值分解, 并计算分解后用于表征各待分析文本的特征向量之间的相似度, 保证了计算得出的文本相似度能够较为准确地体现文本之间的主题相似性, 而并非仅仅依赖于文本之间所共同存在的单词来计算相似度, 因此, 提高了文本相似度的计算准确率。并且, 在一次运算过程中, 本申请实施例能够同时计算参照文本分别与多个对比文本之间的相似度, 故也提高了文本的比对效率。

作为本申请另一实施例, 如图3所示, 在上述S104之后, 还包括:

S108: 获取超参数值, 所述超参数值根据用户预设的矩阵目标维度来确定。

S109: 基于所述超参数值, 分别对各个分解矩阵进行降维处理, 以使所述每一所述分解矩阵的特征维度变更为所述超参数值; 其中, 所述分解矩阵包括所述单词特征矩阵、所述文档特征矩阵以及所述特征向量比重矩阵。

5       本申请实施例中, 在对文本集特征矩阵进行奇异值分解, 得到单词特征矩阵、文档特征矩阵和特征向量比重矩阵等多个分解矩阵后, 由于各个分解矩阵的特征维度较大, 故需要对其特征维度进行降维处理。

具体地, 分别对单词特征矩阵、文档特征矩阵以及特征向量比重矩阵中的多个矩阵元素执行删除操作, 以实现各个分解矩阵的降维处理。在获取预设的超参数值(上述超参数值表示用户所预先设置的矩阵目标)后, 判断降维后每一分解矩阵的特征维度与超参数值是否相同。若存在特征维度与超参数值不相同的分解矩阵, 则将该分解矩阵执行进一步的降维处理, 直至其特征维度与超参数值相同。

作为本申请的一个实施例, 图 4 示出了本申请实施例提供的文本相似度的获取方法 S109 的具体实现流程, 详述如下:

15       S1091: 在每一所述分解矩阵中, 根据矩阵元素的大小顺序, 对该分解矩阵所包含的各个所述矩阵元素进行排序处理, 并读取排序后所述分解矩阵中所存在的各个矩阵列; 每一所述矩阵列包含预设数量的所述矩阵元素。

本申请实施例中, 单词特征矩阵、文档特征矩阵以及特征向量比重矩阵所包含的矩阵元素均不相同。在每一分解矩阵中, 根据各个矩阵元素的数值大小, 对该分解矩阵所包含的各个矩阵元素进行重新排序。由于分解矩阵的行列数由词袋模型所包含的分词数目以及比对本申请的数目来确定, 故矩阵元素重新排序后所得到的分解矩阵的行列数以及矩阵元素数目均保持不变。

S1092: 提取排序在前的多个所述矩阵列, 并将提取得到的多个所述矩阵列组合为更新后的所述分解矩阵; 其中, 所述矩阵列的提取数目与所述超参数值相等。

25       本申请实施例中, 根据每一矩阵列的列序号, 从上述重新排序后的分解矩阵中提取排序在前的 N 个矩阵列。其中, N 为上述预先获取到的超参数值, N 为大于零的整数。

将提取得到的 N 个矩阵列进行依次排列, 以组合得到另一分解矩阵后, 删除原本包含该 N 个矩阵列的分解矩阵, 由此实现分解矩阵的更新及降维处理。

值得注意的是, 本申请实施例中的分解矩阵包括上述单词特征矩阵、文档特征矩阵以及特征向量比重矩阵中的一个或多个, 在此不做限定。

特别地, 当上述分解矩阵为特征向量比重矩阵时, 如图 5 所示, 上述步骤 S1091 具体

包括：

S10911：识别所述特征向量比重矩阵中的各个奇异值，所述奇异值为所述特征向量比重矩阵中的非零值矩阵元素。

5 S10912：基于所述奇异值的数值大小，依次对各个所述奇异值所属的所述矩阵列进行排序，并读取排序后所得到的各个所述矩阵列；每一所述矩阵列包含预设数量的所述矩阵元素。

在对文本集特征矩阵进行奇异值分解后所得到的特征向量比重矩阵中，每一矩阵列将包含有一非零值的矩阵元素，则该矩阵元素为特征向量比值矩阵的奇异值。本申请实施例中，检测出特征向量比值矩阵中的各个奇异值后，依照各奇异值的数值大小，对各个奇异值所属的矩阵列进行排序处理，以使包含数值较大的奇异值的矩阵列能够排列于包含数值较小的奇  
10 异值的矩阵列之前。在对各个矩阵列的排序进行调整后，将当前的特征向量比值矩阵确定为更新后的特征向量比值矩阵。此时，再基于上述步骤 S1092，提取当前特征向量比值矩阵中排序在前的 N 个矩阵列。

例如，上一实施例所示例的特征向量比值矩阵 S 如下：

$$S = \begin{bmatrix} 4.0989 & 0.0000 & 0.0000 \\ 0.0000 & 2.3616 & 0.0000 \\ 0.0000 & 0.0000 & 1.2737 \end{bmatrix}$$

15 由于该特征向量比值矩阵所包含的奇异值分别为 4.0989、2.3616 以及 1.2737，且每一奇异值的数值依序递减，因此，更新后的特征向量比值矩阵 S 不变，即：将 4.0989 所属的矩阵列排列作为特征向量比值矩阵 S 中的第一个矩阵列，将 2.3616 所属的矩阵列排列作为特征向量比值矩阵 S 中的第二个矩阵列，将 1.2737 所属的矩阵列排列作为特征向量比值矩阵 S 中的第三个矩阵列。

20 优选地，在更新后的特征向量比值矩阵 S 中，对每相邻的两个矩阵列，计算排序在前的奇异值与排序在后的奇异值的比值。若该比值大于预设阈值，则获取其中排序较前的矩阵列的序列号，并将该序列号更新为当前时刻的超参数值。

例如，若第二列矩阵列所包含的奇异值为 4.0989，第三列矩阵列所包含的奇异值为 0.0589，且预设阈值为 50，则上述两个奇异值的比值 69.59 大于 50，故将当前的超参数值确  
25 定为排序在前的矩阵列的序列号，即，将当前的超参数值确定为 2，由此实现了超参数值的自动确定及更新。

本申请实施例中，各个奇异值沿着特征向量比值矩阵的对角线由大到小依次排列，且数值下降速度通常较快，由于数值较小的奇异值较难体现比对文本的主题特征，因此，通过将数值较大的 N 个奇异值所在的矩阵列进行提取，将其他矩阵列进行删除，在降低了特征向

量比值矩阵的矩阵维度的同时，保证了更新后的特征向量比值矩阵能够维持比对文本的主题特征，故提高了文本相似度的计算准确性。

作为本申请的一个实施例，图 6 示出了本申请实施例提供的文本相似度的获取方法 S102 的具体实现流程，详述如下：

S1021：分别对每一待分析文本进行分词处理，以得到每一所述待分析文本所对应的多个分词。

S1022：根据预设的停用词列表，检测各个所述分词中是否存在停用词。

S1023：若存在与所述停用词匹配的所述分词，则将该分词进行删除，并将剩余的各个所述分词存入词袋模型。

在待分析文本中，不具备实际语义的字符或词组为上述停用词，即，对待分析文本所表达主题特征的影响程度相对较小的词语为停用词。停用词包括但不限于语气助词、副词、介词以及连接词等。

本申请实施例中，在加载停用词列表之前，获取多个预设的字词，并在预设的时间窗口内，检测每一字符在各网页文章中所出现的总次数。对任一字符，若该字符在各网页文章中所出现的总次数大于第一阈值，且在每一网页文章中的出现概率大于第二阈值，则判断该字符为停用词。此时，将该字符添加至停用词列表。

通过加载停用词列表，获取预存储于停用词列表中的各个停用词。对待分析文本所对应的每一分词，检测其与停用词列表中的停用词是否相同。若检测结果为是，则将该分词进行删除处理，以减少待分析文本所对应的分词的总数。仅当待分词文本所对应的每一分词均与停用词列表中的停用词不相同，才将当前的各个分词存入词袋模型。

本申请实施例中，通过利用预设的停用词列表来检测各个分词中所存在停用词，并将其中的停用词删除后，将剩余的其他分词存入词袋模型，能够减少词袋模型的分词总数，保证了词袋模型中的各个分词均能够表征待分析文本的主题特征，因而降低了文本相似度的计算复杂度，提高了文本相似度的获取效率。

应理解，上述实施例中各步骤的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不对本申请实施例的实施过程构成任何限定。

对应于本申请实施例所提供的文本相似度的获取方法，图 7 示出了本申请实施例提供的文本相似度的获取装置的结构框图。为了便于说明，仅示出了与本实施例相关的部分。

参照图 7，该装置包括：

第一获取单元 71，用于获取预设的待分析文本，所述待分析文本包括参照文本以及多

个比对文本。

分词单元 72，用于对所述待分析文本进行分词处理，以得到每一所述待分析文本所分别对应的多个分词后，将所述分词存入词袋模型。

5 第二获取单元 73，用于对每一所述待分析文本所对应的的每一所述分词，获取该分词在所述词袋模型中的词频逆文本频率 TF-IDF 信息。

生成单元 74，用于基于各个所述待分析文本所关联的所述 TF-IDF 信息，分别生成与所述多个比对文本对应的文本集特征矩阵以及与所述参照文本对应的文本向量；其中，所述文本集特征矩阵包含多个第一特征向量，每一所述第一特征向量与一所述比对文本匹配。

10 分解单元 75，用于对所述文本集特征矩阵进行奇异值分解，得到单词特征矩阵、文档特征矩阵和特征向量比重矩阵。

映射单元 76，用于根据所述单词特征矩阵以及所述特征向量比重矩阵，对所述文本向量进行反向映射处理，得到与所述参照文本匹配的第二特征向量。

计算单元 77，用于分别计算每一所述第二特征向量与所述第一特征向量的相似度，并将计算结果输出为所述参照文本与该第二特征向量所匹配的所述比对文本之间的相似度。

15 可选地，上述文本相似度的获取装置还包括：

第三获取单元，用于获取超参数值，所述超参数值根据用户预设的矩阵目标维度来确。

降维单元，用于基于所述超参数值，分别对各个分解矩阵进行降维处理，以使所述每一所述分解矩阵的特征维度变更为所述超参数值。

20 其中，所述分解矩阵包括所述单词特征矩阵、所述文档特征矩阵以及所述特征向量比重矩阵。

可选地，所述降维单元包括：

排序子单元，用于在每一所述分解矩阵中，根据矩阵元素的大小顺序，对该分解矩阵所包含的各个所述矩阵元素进行排序处理，并读取排序后所述分解矩阵中所存在的各个矩阵列；每一所述矩阵列包含预设数量的所述矩阵元素。

25 提取子单元，用于提取排序在前的多个所述矩阵列，并将提取得到的多个所述矩阵列组合为更新后的所述分解矩阵；其中，所述矩阵列的提取数目与所述超参数值相等。

可选地，所述排序子单元具体用于：

若所述分解矩阵为特征向量比重矩阵，则识别所述特征向量比重矩阵中的各个奇异值，所述奇异值为所述特征向量比重矩阵中的非零值矩阵元素；

30 基于所述奇异值的数值大小，依次对各个所述奇异值所属的所述矩阵列进行排序，并读取排序后所得到的各个所述矩阵列；每一所述矩阵列包含预设数量的所述矩阵元素。

可选地，所述分词单元 72 包括：

分词子单元，用于分别对每一待分析文本进行分词处理，以得到每一所述待分析文本所对应的多个分词。

检测子单元，用于根据预设的停用词列表，检测各个所述分词中是否存在停用词。

5 删除子单元，用于若存在与所述停用词匹配的所述分词，则将该分词进行删除，并将剩余的各个所述分词存入词袋模型。

图 8 是本申请一实施例提供的终端设备的示意图。如图 8 所示，该实施例的终端设备 8 包括：处理器 80 以及存储器 81，所述存储器 81 中存储有可在所述处理器 80 上运行的计算机可读指令 82，例如文本相似度的获取程序。所述处理器 80 执行所述计算机可读指令 82 10 时实现上述各个文本相似度的获取方法实施例中的步骤，例如图 1 所示的步骤 101 至 107。或者，所述处理器 80 执行所述计算机可读指令 82 时实现上述各装置实施例中各模块/单元的功能，例如图 7 所示单元 71 至 77 的功能。

示例性的，所述计算机可读指令 82 可以被分割成一个或多个模块/单元，所述一个或者多个模块/单元被存储在所述存储器 81 中，并由所述处理器 80 执行，以完成本申请。所述 15 一个或多个模块/单元可以是能够完成特定功能的一系列计算机可读指令段，该指令段用于描述所述计算机可读指令 82 在所述终端设备 8 中的执行过程。

所述终端设备 8 可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。所述终端设备可包括，但不仅限于处理器 80 和存储器 81。本领域技术人员可以理解，图 8 仅仅是终端设备 8 的示例，并不构成对终端设备 8 的限定，可以包括比图示更多或更少的部 20 件，或者组合某些部件，或者不同的部件，例如所述终端设备还可以包括输入输出设备、网络接入设备、总线等。

所称处理器 80 可以是中央处理单元(Central Processing Unit, CPU)，还可以是其他通用处理器、数字信号处理器 (Digital Signal Processor, DSP)、专用集成电路 (Application Specific Integrated Circuit, ASIC)、现成可编程门阵列 (Field-Programmable Gate Array, FPGA) 或者 25 其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

所述存储器 81 可以是所述终端设备 8 的内部存储单元，例如终端设备 8 的硬盘或内存。所述存储器 81 也可以是所述终端设备 8 的外部存储设备，例如所述终端设备 8 上配备的插接式硬盘，智能存储卡 (Smart Media Card, SMC)，安全数字 (Secure Digital, SD) 卡， 30 闪存卡 (Flash Card) 等。进一步地，所述存储器 81 还可以既包括所述终端设备 8 的内部存储单元也包括外部存储设备。所述存储器 81 用于存储所述计算机可读指令以及所述终端设

备所需的其他程序和数据。所述存储器 81 还可以用于暂时地存储已经输出或者将要输出的数据。

另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（Read-Only Memory, ROM）、随机存取存储器（Random Access Memory, RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

以上所述，以上实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

## 权利要求书

1. 一种文本相似度的获取方法，其特征在于，包括：

获取预设的待分析文本，所述待分析文本包括参照文本以及多个比对文本；

对所述待分析文本进行分词处理，以得到每一所述待分析文本所分别对应的多个分词  
5 后，将所述分词存入词袋模型；

对每一所述待分析文本所对应的的每一所述分词，获取该分词在所述词袋模型中的词频  
逆文本频率 TF-IDF 信息；

基于各个所述待分析文本所关联的所述 TF-IDF 信息，分别生成与所述多个比对文本对  
应的文本集特征矩阵以及与所述参照文本对应的文本向量；其中，所述文本集特征矩阵包含  
10 多个第一特征向量，每一所述第一特征向量与一所述比对文本匹配；

对所述文本集特征矩阵进行奇异值分解，得到单词特征矩阵、文档特征矩阵和特征向  
量比重矩阵；

根据所述单词特征矩阵以及所述特征向量比重矩阵，对所述文本向量进行反向映射处  
理，得到与所述参照文本匹配的第二特征向量；

15 分别计算每一所述第二特征向量与所述第一特征向量的相似度，并将计算结果输出为  
所述参照文本与该第二特征向量所匹配的所述比对文本之间的相似度。

2. 如权利要求 1 所述的文本相似度的获取方法，其特征在于，在所述对所述文本集特  
征矩阵进行奇异值分解，得到单词特征矩阵、文档特征矩阵和特征向量比重矩阵之后，还包  
括：

20 获取超参数值，所述超参数值根据用户预设的矩阵目标维度来确定；

基于所述超参数值，分别对各个分解矩阵进行降维处理，以使所述每一所述分解矩阵  
的特征维度变更为所述超参数值；

其中，所述分解矩阵包括所述单词特征矩阵、所述文档特征矩阵以及所述特征向量比  
重矩阵。

25 3. 如权利要求 2 所述的文本相似度的获取方法，其特征在于，所述基于所述超参数值，  
分别对各个分解矩阵进行降维处理，以使所述每一所述分解矩阵的特征维度变更为所述超参  
数值，包括：

在每一所述分解矩阵中，根据矩阵元素的大小顺序，对该分解矩阵所包含的各个所述  
矩阵元素进行排序处理，并读取排序后所述分解矩阵中所存在的各个矩阵列；每一所述矩阵  
30 列包含预设数量的所述矩阵元素；

提取排序在前的多个所述矩阵列，并将提取得到的多个所述矩阵列组合为更新后的所述分解矩阵；其中，所述矩阵列的提取数目与所述超参数值相等。

4. 如权利要求 3 所述的文本相似度的获取方法，其特征在于，所述在每一所述分解矩阵中，根据矩阵元素的大小顺序，对该分解矩阵所包含的各个所述矩阵元素进行排序处理，并读取排序后所述分解矩阵中所存在的各个矩阵列，包括：

若所述分解矩阵为特征向量比重矩阵，则识别所述特征向量比重矩阵中的各个奇异值，所述奇异值为所述特征向量比重矩阵中的非零值矩阵元素；

基于所述奇异值的数值大小，依次对各个所述奇异值所属的所述矩阵列进行排序，并读取排序后所得到的各个所述矩阵列；每一所述矩阵列包含预设数量的所述矩阵元素。

5. 如权利要求 1 所述的文本相似度的获取方法，其特征在于，所述对所述待分析文本进行分词处理，以得到每一所述待分析文本所分别对应的多个分词后，将所述分词存入词袋模型，包括：

分别对每一待分析文本进行分词处理，以得到每一所述待分析文本所对应的多个分词；根据预设的停用词列表，检测各个所述分词中是否存在停用词；

若存在与所述停用词匹配的所述分词，则将该分词进行删除，并将剩余的各个所述分词存入词袋模型。

6. 一种文本相似度的获取装置，其特征在于，包括：

第一获取单元，用于获取预设的待分析文本，所述待分析文本包括参照文本以及多个比对文本；

分词单元，用于对所述待分析文本进行分词处理，以得到每一所述待分析文本所分别对应的多个分词后，将所述分词存入词袋模型；

第二获取单元，用于对每一所述待分析文本所对应的的每一所述分词，获取该分词在所述词袋模型中的词频逆文本频率 TF-IDF 信息；

生成单元，用于基于各个所述待分析文本所关联的所述 TF-IDF 信息，分别生成与多个比对文本对应的文本集特征矩阵以及与所述参照文本对应的文本向量；其中，所述文本集特征矩阵包含多个第一特征向量，每一所述第一特征向量与一所述比对文本匹配；

分解单元，用于对所述文本集特征矩阵进行奇异值分解，得到单词特征矩阵、文档特征矩阵和特征向量比重矩阵；

映射单元，用于根据所述单词特征矩阵以及所述特征向量比重矩阵，对所述文本向量进行反向映射处理，得到与所述参照文本匹配的第二特征向量；

计算单元，用于分别计算每一所述第二特征向量与所述第一特征向量的相似度，并将

计算结果输出为所述参照文本与该第二特征向量所匹配的所述比对文本之间的相似度。

7. 根据权利要求 6 所述的文本相似度的获取装置，其特征在于，还包括：

第三获取单元，用于获取超参数值，所述超参数值根据用户预设的矩阵目标维度来确；

降维单元，用于基于所述超参数值，分别对各个分解矩阵进行降维处理，以使所述每一所述分解矩阵的特征维度变更为所述超参数值；

其中，所述分解矩阵包括所述单词特征矩阵、所述文档特征矩阵以及所述特征向量比重矩阵。

8. 根据权利要求 7 所述的文本相似度的获取装置，其特征在于，所述降维单元包括：

排序子单元，用于在每一所述分解矩阵中，根据矩阵元素的大小顺序，对该分解矩阵所包含的各个所述矩阵元素进行排序处理，并读取排序后所述分解矩阵中所存在的各个矩阵列；每一所述矩阵列包含预设数量的所述矩阵元素；

提取子单元，用于提取排序在前的多个所述矩阵列，并将提取得到的多个所述矩阵列组合为更新后的所述分解矩阵；其中，所述矩阵列的提取数目与所述超参数值相等。

9. 根据权利要求 8 所述的文本相似度的获取装置，其特征在于，所述排序子单元具体用于：

若所述分解矩阵为特征向量比重矩阵，则识别所述特征向量比重矩阵中的各个奇异值，所述奇异值为所述特征向量比重矩阵中的非零值矩阵元素；

基于所述奇异值的数值大小，依次对各个所述奇异值所属的所述矩阵列进行排序，并读取排序后所得到的各个所述矩阵列；每一所述矩阵列包含预设数量的所述矩阵元素。

10. 根据权利要求 6 所述的文本相似度的获取装置，其特征在于，所述分词单元包括：分词子单元，用于分别对每一待分析文本进行分词处理，以得到每一所述待分析文本所对应的多个分词；

检测子单元，用于根据预设的停用词列表，检测各个所述分词中是否存在停用词；

删除子单元，用于若存在与所述停用词匹配的所述分词，则将该分词进行删除，并将剩余的各个所述分词存入词袋模型。

11. 一种终端设备，其特征在于，包括存储器以及处理器，所述存储器中存储有可在所述处理器上运行的计算机可读指令，所述处理器执行所述计算机可读指令时实现如下步骤：

获取预设的待分析文本，所述待分析文本包括参照文本以及多个比对文本；

对所述待分析文本进行分词处理，以得到每一所述待分析文本所分别对应的多个分词后，将所述分词存入词袋模型；

对每一所述待分析文本所对应的每一所述分词，获取该分词在所述词袋模型中的词频

逆文本频率 TF-IDF 信息;

基于各个所述待分析文本所关联的所述 TF-IDF 信息, 分别生成与所述多个比对文本对应的文本集特征矩阵以及与所述参照文本对应的文本向量; 其中, 所述文本集特征矩阵包含多个第一特征向量, 每一所述第一特征向量与一所述比对文本匹配;

- 5        对所述文本集特征矩阵进行奇异值分解, 得到单词特征矩阵、文档特征矩阵和特征向量比重矩阵;

根据所述单词特征矩阵以及所述特征向量比重矩阵, 对所述文本向量进行反向映射处理, 得到与所述参照文本匹配的第二特征向量;

- 10       分别计算每一所述第二特征向量与所述第一特征向量的相似度, 并将计算结果输出为所述参照文本与该第二特征向量所匹配的所述比对文本之间的相似度。

12. 根据权利要求 11 所述的终端设备, 其特征在于, 所述处理器执行所述计算机可读指令时还实现如下步骤:

获取超参数值, 所述超参数值根据用户预设的矩阵目标维度来确定;

- 15       基于所述超参数值, 分别对各个分解矩阵进行降维处理, 以使所述每一所述分解矩阵的特征维度变更为所述超参数值;

其中, 所述分解矩阵包括所述单词特征矩阵、所述文档特征矩阵以及所述特征向量比重矩阵。

- 20       13. 根据权利要求 12 所述的终端设备, 其特征在于, 所述基于所述超参数值, 分别对各个分解矩阵进行降维处理, 以使所述每一所述分解矩阵的特征维度变更为所述超参数值, 包括:

在每一所述分解矩阵中, 根据矩阵元素的大小顺序, 对该分解矩阵所包含的各个所述矩阵元素进行排序处理, 并读取排序后所述分解矩阵中所存在的各个矩阵列; 每一所述矩阵列包含预设数量的所述矩阵元素;

- 25       提取排序在前的多个所述矩阵列, 并将提取得到的多个所述矩阵列组合为更新后的所述分解矩阵; 其中, 所述矩阵列的提取数目与所述超参数值相等。

14. 根据权利要求 13 所述的终端设备, 其特征在于, 所述在每一所述分解矩阵中, 根据矩阵元素的大小顺序, 对该分解矩阵所包含的各个所述矩阵元素进行排序处理, 并读取排序后所述分解矩阵中所存在的各个矩阵列, 包括:

- 30       若所述分解矩阵为特征向量比重矩阵, 则识别所述特征向量比重矩阵中的各个奇异值, 所述奇异值为所述特征向量比重矩阵中的非零值矩阵元素;

基于所述奇异值的数值大小, 依次对各个所述奇异值所属的所述矩阵列进行排序, 并

读取排序后所得到的各个所述矩阵列；每一所述矩阵列包含预设数量的所述矩阵元素。

15. 根据权利要求 11 所述的终端设备，其特征在于，所述对所述待分析文本进行分词处理，以得到每一所述待分析文本所分别对应的多个分词后，将所述分词存入词袋模型，包括：

5        分别对每一待分析文本进行分词处理，以得到每一所述待分析文本所对应的多个分词；  
根据预设的停用词列表，检测各个所述分词中是否存在停用词；

若存在与所述停用词匹配的所述分词，则将该分词进行删除，并将剩余的各个所述分词存入词袋模型。

16. 一种计算机可读存储介质，所述计算机可读存储介质存储有计算机可读指令，其特征在于，所述计算机可读指令被至少一个处理器执行时实现如下步骤：

获取预设的待分析文本，所述待分析文本包括参照文本以及多个比对文本；

对所述待分析文本进行分词处理，以得到每一所述待分析文本所分别对应的多个分词后，将所述分词存入词袋模型；

对每一所述待分析文本所对应的的每一所述分词，获取该分词在所述词袋模型中的词频  
15 逆文本频率 TF-IDF 信息；

基于各个所述待分析文本所关联的所述 TF-IDF 信息，分别生成与所述多个比对文本对应的文本集特征矩阵以及与所述参照文本对应的文本向量；其中，所述文本集特征矩阵包含多个第一特征向量，每一所述第一特征向量与一所述比对文本匹配；

对所述文本集特征矩阵进行奇异值分解，得到单词特征矩阵、文档特征矩阵和特征向量  
20 量比重矩阵；

根据所述单词特征矩阵以及所述特征向量比重矩阵，对所述文本向量进行反向映射处理，得到与所述参照文本匹配的第二特征向量；

分别计算每一所述第二特征向量与所述第一特征向量的相似度，并将计算结果输出为所述参照文本与该第二特征向量所匹配的所述比对文本之间的相似度。

25        17. 根据权利要求 16 所述的计算机可读存储介质，其特征在于，所述计算机可读指令被至少一个处理器执行时还实现如下步骤：

获取超参数值，所述超参数值根据用户预设的矩阵目标维度来确定；

基于所述超参数值，分别对各个分解矩阵进行降维处理，以使所述每一所述分解矩阵的特征维度变更为所述超参数值；

30        其中，所述分解矩阵包括所述单词特征矩阵、所述文档特征矩阵以及所述特征向量比重矩阵。

18. 根据权利要求 17 所述的计算机可读存储介质，其特征在于，所述基于所述超参数值，分别对各个分解矩阵进行降维处理，以使所述每一所述分解矩阵的特征维度变更为所述超参数值，包括：

5 在每一所述分解矩阵中，根据矩阵元素的大小顺序，对该分解矩阵所包含的各个所述矩阵元素进行排序处理，并读取排序后所述分解矩阵中所存在的各个矩阵列；每一所述矩阵列包含预设数量的所述矩阵元素；

提取排序在前的多个所述矩阵列，并将提取得到的多个所述矩阵列组合为更新后的所述分解矩阵；其中，所述矩阵列的提取数目与所述超参数值相等。

10 19. 根据权利要求 18 所述的计算机可读存储介质，其特征在于，所述在每一所述分解矩阵中，根据矩阵元素的大小顺序，对该分解矩阵所包含的各个所述矩阵元素进行排序处理，并读取排序后所述分解矩阵中所存在的各个矩阵列，包括：

若所述分解矩阵为特征向量比重矩阵，则识别所述特征向量比重矩阵中的各个奇异值，所述奇异值为所述特征向量比重矩阵中的非零值矩阵元素；

15 基于所述奇异值的数值大小，依次对各个所述奇异值所属的所述矩阵列进行排序，并读取排序后所得到的各个所述矩阵列；每一所述矩阵列包含预设数量的所述矩阵元素。

20 20. 根据权利要求 16 所述的计算机可读存储介质，其特征在于，所述对所述待分析文本进行分词处理，以得到每一所述待分析文本所分别对应的多个分词后，将所述分词存入词袋模型，包括：

分别对每一待分析文本进行分词处理，以得到每一所述待分析文本所对应的多个分词；

20 根据预设的停用词列表，检测各个所述分词中是否存在停用词；

若存在与所述停用词匹配的所述分词，则将该分词进行删除，并将剩余的各个所述分词存入词袋模型。

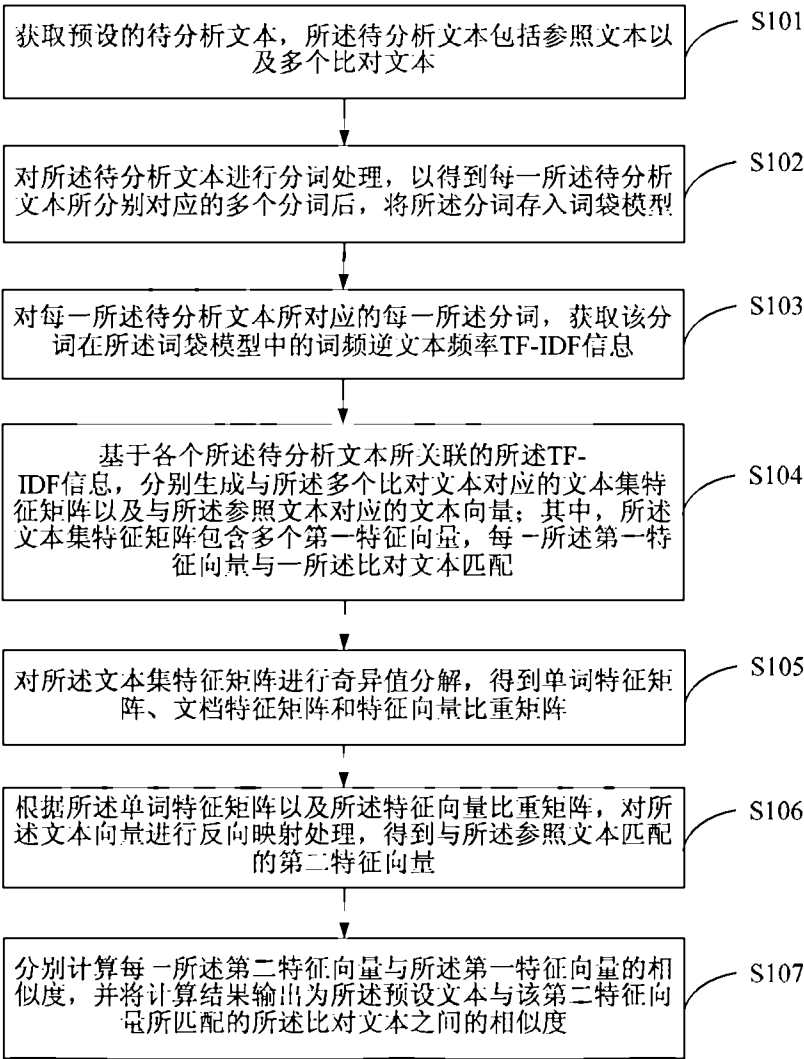


图 1

| 单词       | d1 | d2 | d3 | q |
|----------|----|----|----|---|
| a        | 1  | 1  | 1  | 0 |
| arrived  | 0  | 1  | 1  | 0 |
| damaged  | 1  | 0  | 0  | 0 |
| delivery | 0  | 1  | 0  | 0 |
| fire     | 1  | 0  | 0  | 0 |
| gold     | 1  | 0  | 1  | 1 |
| in       | 1  | 1  | 1  | 0 |
| of       | 1  | 1  | 1  | 0 |
| shipment | 1  | 0  | 1  | 0 |
| silver   | 0  | 2  | 0  | 1 |
| truck    | 0  | 1  | 1  | 1 |

图 2

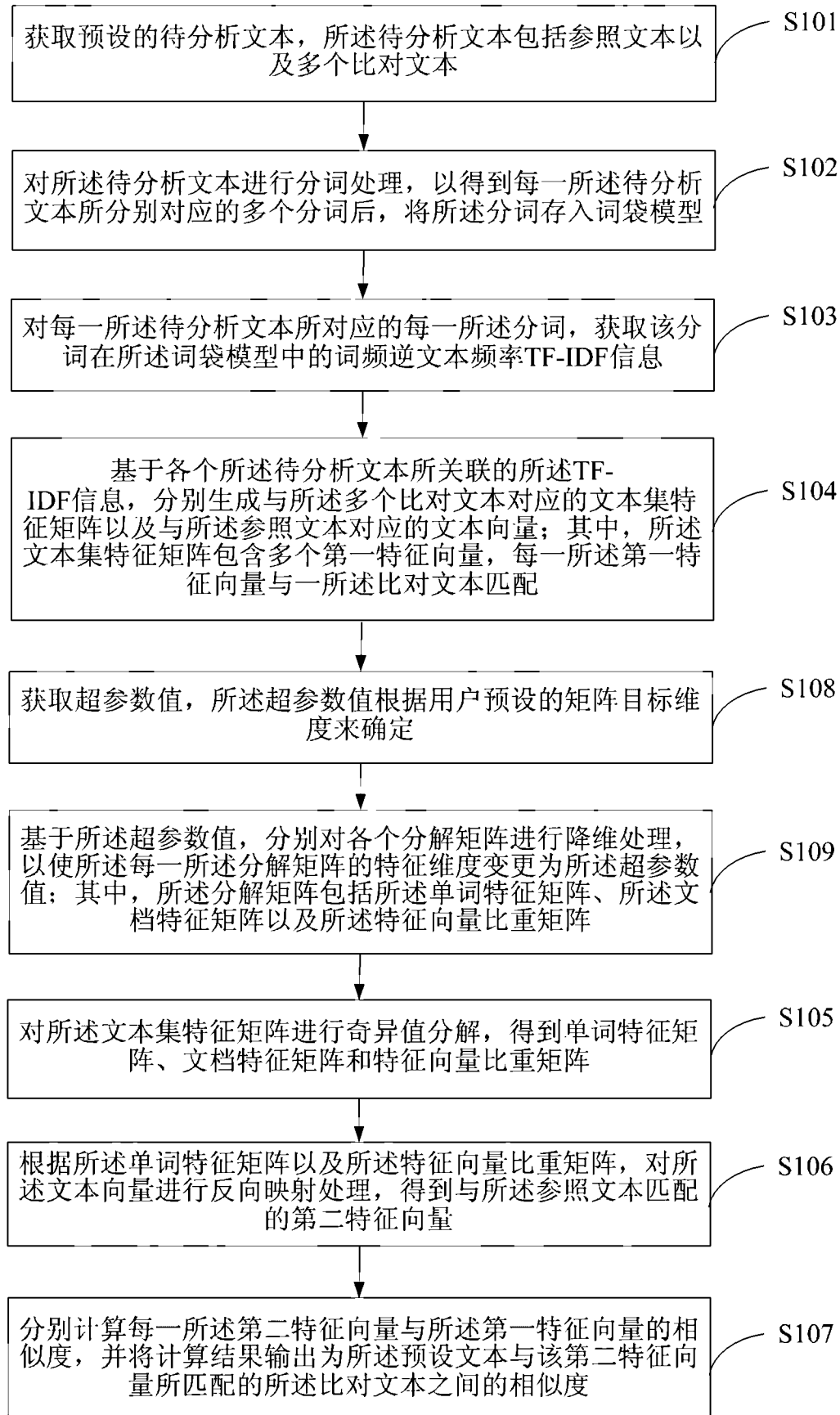


图 3

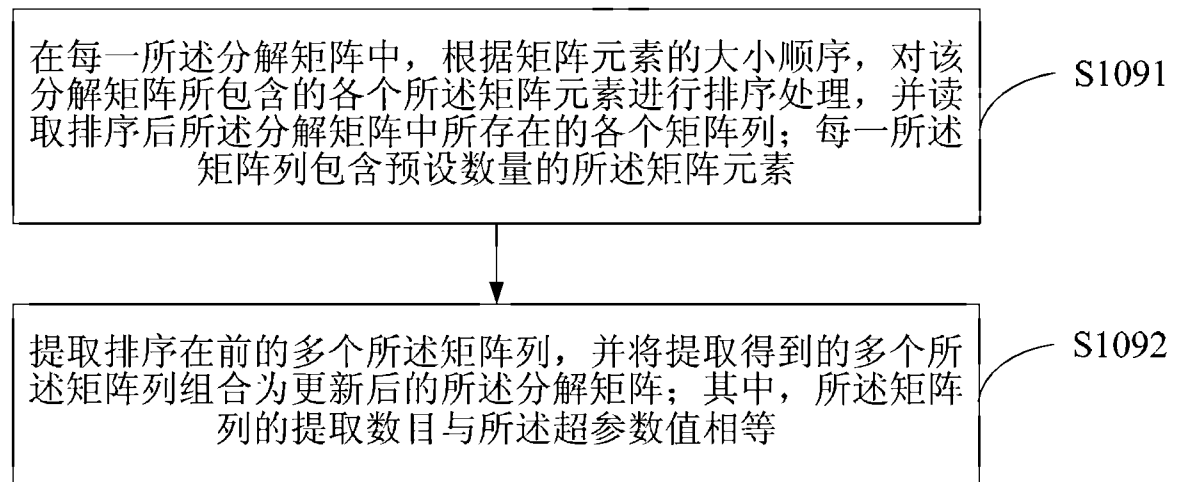


图 4

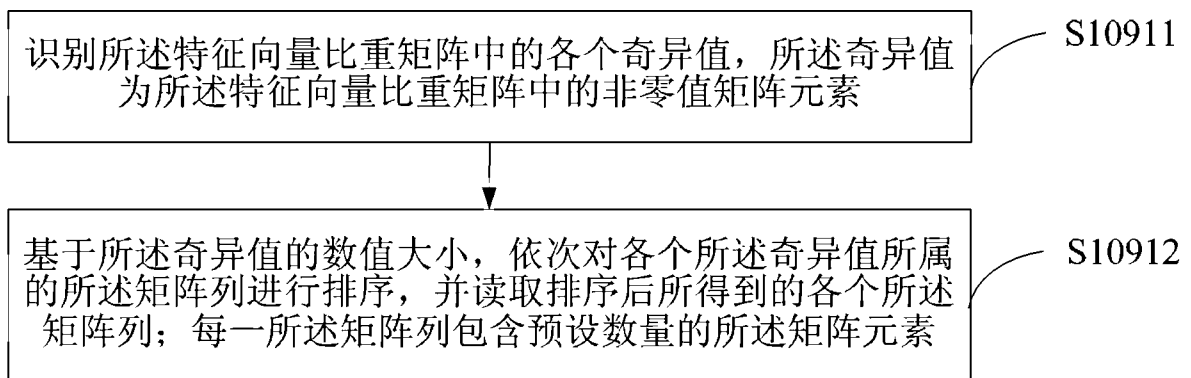


图 5

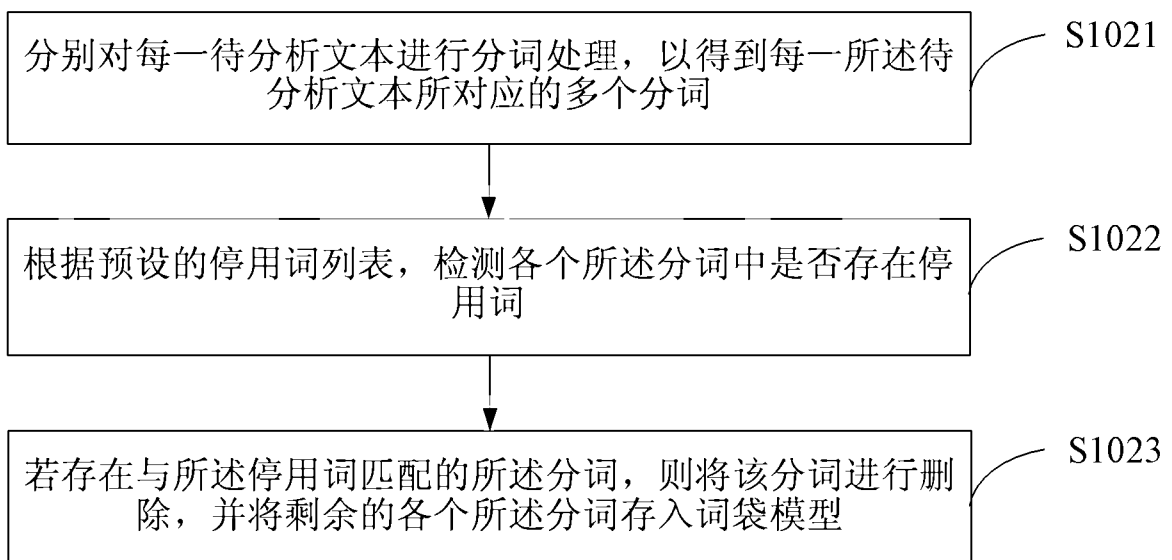


图 6

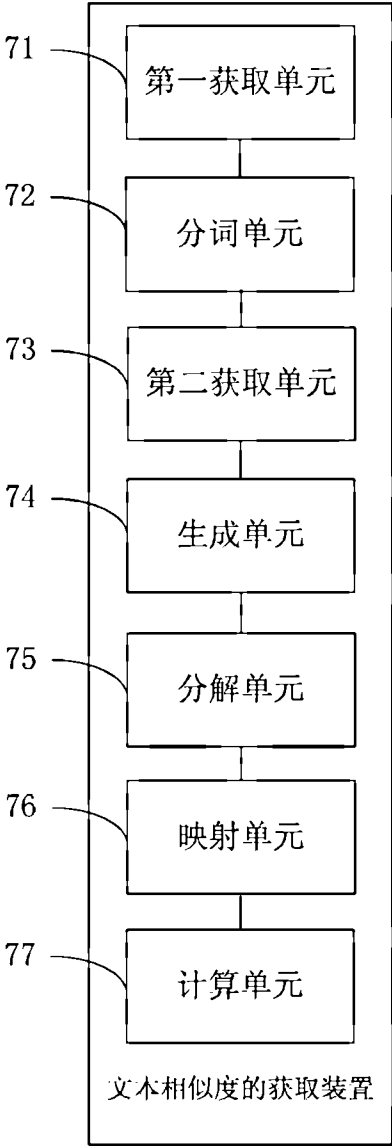


图 7

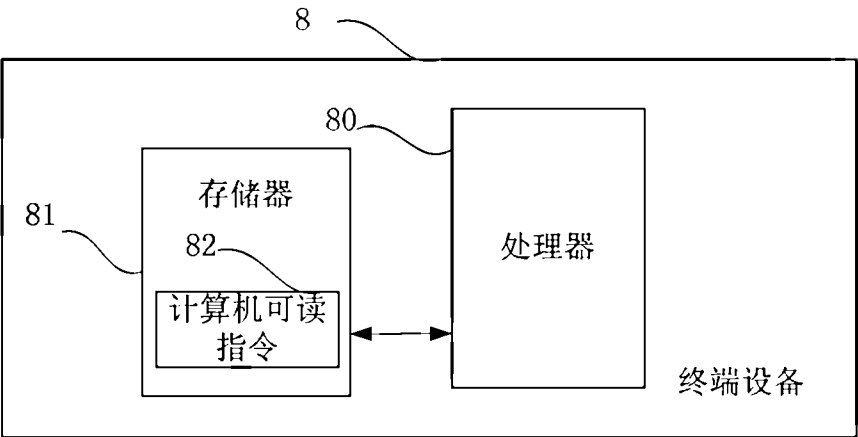


图 8

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/097497

**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)

Data Bases: DWPI, CNABS, SIPOABS, CNKI; Key Terms: 文本, 相似度, 近似度, 矩阵, 向量, 词频逆文本频率, 分词, 模型, 权重; text, similarity, matrix, vector, TF-IDF, segmentation, model, weight

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                              | Relevant to claim No. |
|-----------|-----------------------------------------------------------------------------------------------------------------|-----------------------|
| A         | CN 101251841 A (EAST CHINA NORMAL UNIVERSITY) 27 August 2008 (2008-08-27)<br>entire document                    | 1-20                  |
| A         | CN 104102626 A (XIAMEN TUIITE INFORMATION TECHNOLOGY CO., LTD.) 15 October 2014 (2014-10-15)<br>entire document | 1-20                  |
| A         | CN 107391727 A (BEIHANG UNIVERSITY) 24 November 2017 (2017-11-24)<br>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

“&amp;” document member of the same patent family

Date of the actual completion of the international search

18 February 2019

Date of mailing of the international search report

25 February 2019

Name and mailing address of the ISA/CN

State Intellectual Property Office of the P. R. China (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/CN2018/097497**

| Patent document<br>cited in search report |           |   | Publication date<br>(day/month/year) | Patent family member(s) | Publication date<br>(day/month/year) |
|-------------------------------------------|-----------|---|--------------------------------------|-------------------------|--------------------------------------|
| CN                                        | 101251841 | A | 27 August 2008                       | None                    |                                      |
| CN                                        | 104102626 | A | 15 October 2014                      | None                    |                                      |
| CN                                        | 107391727 | A | 24 November 2017                     | None                    |                                      |

## 国际检索报告

国际申请号

PCT/CN2018/097497

## 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))

数据库: DWPI, CNABS, SIPOABS, CNKI; 关键词: 文本, 相似度, 近似度, 矩阵, 向量, 词频逆文本频率, 分词, 模型, 权重; text, similarity, matrix, vector, TF-IDF, segmentation, model, weight

## 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 101251841 A (华东师范大学) 2008年 8月 27日 (2008 - 08 - 27)<br>全文        | 1-20    |
| A    | CN 104102626 A (厦门推特信息科技有限公司) 2014年 10月 15日 (2014 - 10 - 15)<br>全文 | 1-20    |
| A    | CN 107391727 A (北京航空航天大学) 2017年 11月 24日 (2017 - 11 - 24)<br>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

“&amp;” 同族专利的文件

国际检索实际完成的日期

2019年 2月 18日

国际检索报告邮寄日期

2019年 2月 25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

受权官员

刘渊

传真号 (86-10)62019451

电话号码 62412079

国际检索报告  
关于同族专利的信息

国际申请号

PCT/CN2018/097497

| 检索报告引用的专利文件 |           |   | 公布日<br>(年/月/日) | 同族专利 | 公布日<br>(年/月/日) |
|-------------|-----------|---|----------------|------|----------------|
| CN          | 101251841 | A | 2008年 8月 27日   | 无    |                |
| CN          | 104102626 | A | 2014年 10月 15日  | 无    |                |
| CN          | 107391727 | A | 2017年 11月 24日  | 无    |                |