

(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
2020 年 7 月 2 日 (02.07.2020)



(10) 国际公布号
WO 2020/132933 A1

(51) 国际专利分类号:

G06F 16/335 (2019.01)

广东省深圳市南山区学苑大道1001号南山智园C1栋16、22楼, Guangdong 518000 (CN)。

(21) 国际申请号:

PCT/CN2018/123847

(22) 国际申请日:

2018 年 12 月 26 日 (26.12.2018)

(25) 申请语言:

中文

(26) 公布语言:

中文

(30) 优先权:

201811594130.7 2018年12月25日 (25.12.2018) CN

(71) 申请人: 深圳市优必选科技有限公司(**UBTECH ROBOTICS CORP**) [CN/CN]; 中国广东省深圳市南山区学苑大道1001号南山智园C1栋16、22楼, Guangdong 518000 (CN)。

(72) 发明人: 熊友军(**XIONG, Youjun**); 中国广东省深圳市南山区学苑大道1001号南山智园C1栋16、22楼, Guangdong 518000 (CN)。 罗沛鹏(**LUO, Peipeng**); 中国广东省深圳市南山区学苑大道1001号南山智园C1栋16、22楼, Guangdong 518000 (CN)。 廖洪涛(**LIAO, Hongtao**); 中国广

(74) 代理人: 深圳中细软知识产权代理有限公司(**SHENZHEN CIPRUN INTELLECTUAL PROPERTY AGENCY CO., LTD.**); 中国广东省深圳市南山区粤海街道粤兴四道1号中山大学深圳产学研大楼11楼, 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, 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。

(54) **Title:** SHORT TEXT FILTERING METHOD AND APPARATUS, MEDIUM AND COMPUTER DEVICE

(54) 发明名称: 短文本过滤方法、装置、介质及计算机设备

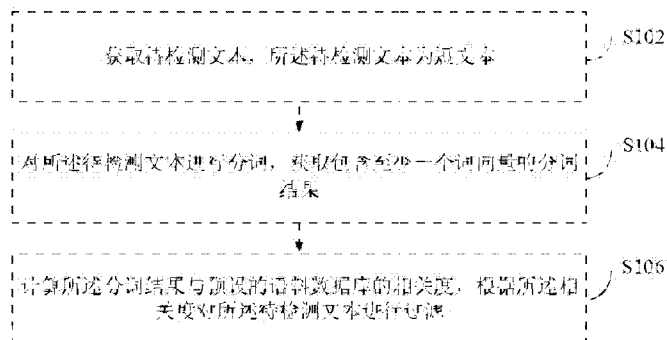


图 1

S102 Acquire a text to be detected, wherein the text to be detected is a short text
S104 Perform word segmentation on the text to be detected, so as to acquire a word segmentation result including at least one word vector
S106 Calculate the relevancy between the word segmentation result and a pre-set corpus database, and filter the text to be detected according to the relevancy

(57) **Abstract:** A short text filtering method and apparatus, a medium and a computer device. The method comprises: acquiring a text to be detected, wherein the text to be detected is a short text (S102); performing word segmentation on the text to be detected, so as to acquire a word segmentation result including at least one word vector (S104); and calculating the relevancy between the word segmentation result and a pre-set corpus database, and filtering the text to be detected according to the relevancy (S106). By means of the short text filtering method and apparatus, the medium and the computer device, the accuracy of short text filtering can be improved.

(57) **摘要:** 一种短文本过滤方法、装置、介质和计算机设备，包括：获取待检测文本，所述待检测文本为短文本 (S102)；对所述待检测文本进行分词，获取包含至少一个词向量的分词结果 (S104)；计算所述分词结果与预设的语料数据库的相关度，根据所述相关度对所述待检测文本进行过滤 (S106)。采用上述短文本过滤方法、装置、介质和计算机设备，可以提高短文本过滤的准确率。

(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))。

短文本过滤方法、装置、介质及计算机设备

技术领域

5 本发明涉及机器学习技术领域，尤其涉及一种短文本过滤方法、装置、计算机设备及存储介质。

背景技术

10 随着深度学习技术的发展和计算机性能的提升，深度学习技术广泛应用于各个领域，尤其是在机器人领域中对机器人进行分类训练或模型训练。然而，在模型训练的过程中，部分与训练任务无关的语料也会进入到模型中，这就使得模型出现误判。

也就是说，深度学习文本、短文本分类模型对未知文本的预测存在准确性不足的问题。

15 发明内容

基于此，有必要针对上述问题，提出一种识别率高的短文本过滤方法、装置和计算机设备。

在本发明的第一方面，提出了一种短文本过滤方法，所述方法包括：

获取待检测文本，所述待检测文本为短文本；

20 对所述待检测文本进行分词，获取包含至少一个词向量的分词结果；

计算所述分词结果与预设的语料数据库的相关度，根据所述相关度对所述待检测文本进行过滤。

在本发明的第二方面，提出了一种短文本过滤装置，所述装置包括：

待检测文本获取模块，用于获取待检测文本，所述待检测文本为短文本；

25 分词处理模块，用于对所述待检测文本进行分词，获取包含至少一个词向量的分词结果；

文本过滤模块，用于计算所述分词结果与预设的语料数据库的相关度，根据所述相关度对所述待检测文本进行过滤。

在本发明的第三方面，还提出了一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述计算机程序被所述处理器执行时，使得所述处理器执行以下步骤：

获取待检测文本，所述待检测文本为短文本；

5 对所述待检测文本进行分词，获取包含至少一个词向量的分词结果；

计算所述分词结果与预设的语料数据库的相关度，根据所述相关度对所述待检测文本进行过滤。

在本发明的第四方面，还提出了一种计算机可读存储介质，存储有计算机程序，所述计算机程序被处理器执行时，使得所述处理器执行以下步骤：

10 获取待检测文本，所述待检测文本为短文本；

对所述待检测文本进行分词，获取包含至少一个词向量的分词结果；

计算所述分词结果与预设的语料数据库的相关度，根据所述相关度对所述待检测文本进行过滤。

本发明提出了一种短文本过滤方法、装置和计算机设备，在需要对短文本进行进一步的分类、预测等操作之前，首先对短文本进行过滤处理，具体通过对短文本进行分词，然后计算分词结果与预设的语料数据库之间是否是相关的来确定该短文本是否为相关文本，根据是否为相关文本对短文本进行过滤。也就是说，只有通过上述过程被判定为相关文本的短文本才会进入到下一个环节进行分类、预测或其他操作，从而提高了深度学习短文本预测的准确性。

20

附图说明

为了更清楚地说明本发明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发明的一些实施例，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。

25

其中：

图 1 为一个实施例中短文本过滤方法的实现流程示意图；

图 2 为一个实施例中短文本过滤方法的实现流程示意图；

图 3 为一个实施例中短文本过滤方法的实现流程示意图；

图 4 为一个实施例中短文本过滤方法的实现流程示意图；

图 5 为一个实施例中短文本过滤方法的实现流程示意图；

图 6 为一个实施例中短文本过滤装置的结构示意图；

图 7 为一个实施例中计算机设备的结构框图。

5

具体实施方式

下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

在本实施例中，特提出了一种短文本过滤方法，该方法的实现可依赖于计算机程序，该计算机程序可运行于基于冯诺依曼体系的计算机系统之上，该计算机程序可以是对文本、短文本进行过滤的应用程序。该计算机系统可以是运行上述计算机程序的例如智能手机、平板电脑、个人电脑等计算机设备。需要说明的是，在本实施例中，上述短文本过滤方法可以用于对机器人进行训练的过程。

如图1所示，在一个实施例中，提供了一种短文本过滤方法，具体包括如下步骤：

步骤 S102：获取待检测文本，所述待检测文本为短文本。

待检测文本，为一个包含一个或者多个汉字的文本，例如，可以通过语音识别获取的文本，也可以是实时获取的用户直接输入的短文本信息，或者是实时获取的计算机设备当前需要处理的短文本信息，还可以是获取预存储在计算机设备或其他设备中的需要进行过滤处理的短文本信息。其中，短文本是相对于长文本来讲的，短文本句子短，信息量少且重要，也即短文本中包含的信息比较稠密。

例如，在本实施例中，待检测文本为：我要去吃饭。

在一个具体的实施例中，待检测文本可以是对机器人进行训练的训练语料。

在本实施例中，在获取到待检测文本之后，还需要对原始的短文本进行一些处理，比如去除停用词（标点符号）、语气助词、口语词、客套词、动词叠

词等，然后才得到最终的待检测的短文本。

在一个具体的实施例中，上述步骤 S102 之后，还包括：按照预设的文本改写算法对所述待检测文本进行改写，所述改写后的待检测文本中不包含预设的辅助词。其中，预设的辅助词包括但不限于停用词（标点符号）、语气助词、
5 口语词、客套词、动词叠词等。如，“左转啊”、“嗯，左转”、“麻烦左转”、“左转转”改为“左转”。

需要说明的是，在本实施例中，对待检测文本的改写还可以不限于上述辅助词的删除，还可以是其他文本的改写，在本实施例中不做限制。

步骤 S104：对所述待检测文本进行分词，获取包含至少一个词向量的分
10 词结果。

在本实施例中，在对短文本进行过滤的过程中，首先要对短文本进行分词处理，将短文本拆分成若干个字或词的组合。例如，将“我想去北京看看”进行分词处理得到的分词结果为：我，想，去，北京，看，看。

在本实施例中，考虑到短文本的特殊性，对待检测文本尽量按字为单位进行切割，但是，除了预设的关键词数据库中包含的关键词除外。也就是说，在
15 对待检测文本进行分词处理的过程中，是根据预设的关键词数据库对所述待检测文本进行分词处理的，且除了在该关键词数据库中包含的关键词之外，其余的均按字为单位进行切割，也就是说，分词结果包括与预设的关键词数据库中包含的关键词对应的词向量、以及若干个字向量。

例如，“我想去北京看看”中“北京”为数据库里的关键词，因此，进行分
20 词处理得到的分词结果为：我，想，去，北京，看，看。

在本实施例中，对短文本尽量按字进行切割时考虑到了短文本的字数较少，按词进行切分可能导致对应的词向量的数量过少，不利于后续的相似度、相关度的计算；而关键词数据库中的关键词予以保留是为了在短文本的基础上尽量
25 多的保留一些重要信息。

步骤 S106：计算所述分词结果与预设的语料数据库的相关度，根据所述相关度对所述待检测文本进行过滤。

待检测文本是否需要进行过滤是通过待检测文本的分词结果与相应的语料数据库之间进行比对来确定的。例如，在待检测文本对应的分词结果与语料

数据库之间是匹配度情况下,待检测文本为相关文本,不需要进行过滤,反之,待检测文本为不相关文本,需要进行过滤。

在本实施例中,通过预设的算法计算待检测文本对应的分词结果与语料数据库之间的相关度,然后根据该相关度信息来确定待检测文本是否为相关文本,并确定是否进行过滤处理。

在一个具体的实施例中,预设的语料数据库包含多条语料文本;在计算待检测文本与语料数据库之间的相关度时,可以通过计算待检测文本对应的分词结果与每一个预料文本之间的相似度来进行。具体的,上述计算所述分词结果与预设的语料数据库的相关度,还包括:按照预设的相似度计算算法分别计算所述分词结果与每一条语料文本之间的相似度;获取所述相似度计算结果中的最大值为目标相似度,作为所述分词结果与预设的语料数据库的相关度。

具体的,待检测文本分词后的分词结果与分完词的语料数据库中的语料文本逐条进行jaccard相似度计算。具体的,令A、B为两个不同的文本(例如,A为待检测文本对应的分词结果,B为语料数据库中的语料文本),则,jaccard相似度公式为:

$$\text{jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

并且,针对所有的语料文本,记录jaccard相似度最高的分值max_jaccard,并将max_jaccard作为目标相似度,并作为分词结果与预设的语料数据库的相关度。

进一步的,如图2所示,上述根据所述相关度对所述待检测文本进行过滤的具体过程为:判断确定目标相似度是否大于或等于预设的第一阈值;若是,判定所述待检测文本为相关文本;反之,所述待检测文本不为相关文本。

例如,在jaccard相似度的最高分值大于第一阈值(0.7)的情况下,即max_jaccard>0.7,待检测文本是预设语料数据库相关的文本,不需要过滤。

在另一个实施例中,待检测文本与语料数据库之间的相关度的计算不仅需要考虑待检测文本与语料文本之间相似度,还需要考虑其他维度的相关度。

具体的,在一个实施例中,上述计算分词结果与预设的语料数据库的相关度的过程还包括:按照预设的编辑距离计算公式分别计算所述分词结果与每一

条语料文本之间的编辑距离；获取所述编辑距离计算结果中的最大值为目标编辑距离，作为所述分词结果与预设的语料数据库的相关度。

待检测文本对应的分词结果与分完词的语料文本逐条进行编辑距离相似度计算，相似度高于预设的第二阈值则为相关文本。

- 5 具体的，令 T 为两个文本（A、B）中词向量/字向量的总数，令 M 为 A、B 中相同字的个数，编辑距离相似度公式为：

$$\text{edit}(A, B) = \frac{2M}{T}$$

并且，针对所有的语料文本，记录编辑距离的最高分值 $\max_edit$ 作为目标编辑距离，即待检测文本与语料数据库之间的编辑距离。

- 10 进一步的，如图 3 所示，上述根据所述相关度对所述待检测文本进行过滤的具体过程为：判断确定目标编辑距离是否大于或等于预设的第二阈值；若是，判定所述待检测文本为相关文本；反之，所述待检测文本不为相关文本。

例如，在编辑距离 edit 的最高分值 $\max_edit$ 大于第二阈值（0.8）的情况下，即 $\max_edit > 0.8$ ，待检测文本是预设语料数据库相关的文本，不需要过滤。

- 15 在另一个可选的实施例中，在考虑是否需要将待检测文本进行过滤时，不仅需要考虑 jaccard 相似度，还需要考虑编辑距离 edit 。

- 具体地，如图 4 所示，判断确定目标相似度是否大于或等于预设的第一阈值；若是，判定所述待检测文本为相关文本；反之，在所述目标相似度小于所述第一阈值的情况下，需要进一步的考虑编辑距离的大小，即进一步判断所述
20 目标编辑距离是否大于或等于预设的第二阈值；若是，判定所述待检测文本为相关文本，反之，待检测文本为不相关文本，需要进行过滤。

- 在另一个实施例中，还可能存在一种情况，相似度和编辑距离均较小，但是其可能也为相关文本。在这种情况下，还可以考虑在待检测文本中包含某些关键词的情况，可以降低对相似度、编辑距离等相关度的要求，从而避免过多的
25 的相关文本被过滤。

如图 5 所示，在目标相似度和目标编辑距离均小于预设的第一/第二阈值的情况下，进一步判断目标相似度是否大于预设第三阈值、且目标编辑距离是否大于预设的第四阈值，在均是的情况下，判断预设的关键词数据库中是否包

含了与所述待检测文本匹配的关键词,若是,判定所述待检测文本为相关文本,反之,待检测文本为不相关文本,需要进行过滤。例如,可以取第三阈值为0.5,第四阈值为0.5。

例如,在 $\text{max_jaccard}=0.6$, $\text{max_edit}=0.55$ 的情况下, $\text{max_jaccard}<0.7$,
5 且 $\text{max_edit}<0.8$,但是,在该待检测文本中包含了预设的关键词,则进一步判断 $\text{max_jaccard}>0.5$, 且 $\text{max_edit}>0.5$,从而判断待检测文本为相关文本。

在本发明实施例中,对于对话、命令等语料对应的短文本,在进行具体的模型训练或分类之前,首先对短文本进行过滤,从而避免了太多的不相关文本进入到训练模型或分类模型,造成较大的误差。具体的,通过对短文本进行分
10 词并将分词结果与预设的语料数据库进行比对,从而确定相应的短文本是否与语料数据库是相关的,从而判断短文本是否为相关文本,是否需要进行过滤,从而大大的提高,深度学习文本、短文本分类模型对未知文本的预测的准确性。

如图6所示,本发明实施例还提供一种短文本过滤装置。具体的,如图6所示,所述短文本过滤装置包括:

15 待检测文本获取模块102,用于获取待检测文本,所述待检测文本为短文本;

分词处理模块104,用于对所述待检测文本进行分词,获取包含至少一个词向量的分词结果;

20 文本过滤模块106,用于计算所述分词结果与预设的语料数据库的相关度,根据所述相关度对所述待检测文本进行过滤。

上述短文本过滤装置,首先获取待检测的短文本(即待检测文本);同时将所述待检测文本进行分词,得到分词结果,所述分词结果中包含多个词向量;然后针对预设语料数据库中包含的语料文本,计算其与待检测文本对应的分词结果之间的相关度,从而根据相关度计算结果来确定待检测文本是否为相关文
25 本,并据此对待检测文本进行过滤。上述对短文本进行过滤减少了不相关的短文本进入到下一步的深度学习(例如模型训练、分类)环节中,提高了深度学习的准确性。

在其中一个实施例中,所述预设的语料数据库包含多条语料文本;所述文本过滤模块106还用于按照预设的相似度计算算法分别计算所述分词结果与

每一条语料文本之间的相似度；获取所述相似度计算结果中的最大值为目标相似度，作为所述分词结果与预设的语料数据库的相关度。

在其中一个实施例中，所述文本过滤模块 106 还用于按照预设的编辑距离计算公式分别计算所述分词结果与每一条语料文本之间的编辑距离；获取所述编辑距离计算结果中的最大值为目标编辑距离，作为所述分词结果与预设的语料数据库的相关度。

在其中一个实施例中，所述文本过滤模块 106 还用于确定所述目标相似度大于或等于预设的第一阈值；在所述目标相似度大于或等于所述第一阈值的情况下，判定所述待检测文本为相关文本；在所述目标相似度小于所述第一阈值的情况下，执行所述按照预设的编辑距离计算公式分别计算所述分词结果与每一条语料文本之间的编辑距离；确定所述目标编辑距离大于或等于预设的第二阈值；判定所述待检测文本为相关文本。

在其中一个实施例中，所述文本过滤模块 106 还用于在所述目标编辑距离小于所述第二阈值的情况下，确定所述目标相似度大于预设第三阈值、且目标编辑距离大于预设的第四阈值；在所述预设的关键词数据库查找与所述待检测文本匹配的关键词关键词，在查找到的情况下，判定所述待检测文本为相关文本。

在其中一个实施例中，如图 6 所示，上述装置还包括文本改写模块 108，用于按照预设的文本改写算法对所述待检测文本进行改写，所述改写后的待检测文本中不包含预设的辅助词。

在其中一个实施例中，所述分词处理模块 104 还用于根据预设的关键词数据库对所述待检测文本进行分词，所述分词结果包括与所述预设的关键词数据库中包含的关键词对应的词向量以及多个字向量。

图 7 示出了一个实施例中计算机设备的内部结构图。该计算机设备具体可以是服务器。如图 7 所示，该计算机设备包括通过系统总线连接的处理器、存储器和网络接口。其中，存储器包括非易失性存储介质和内存储器。该计算机设备的非易失性存储介质存储有操作系统，还可存储有计算机程序，该计算机程序被处理器执行时，可使得处理器实现短文本过滤方法。该内存储器中也可储存有计算机程序，该计算机程序被处理器执行时，可使得处理器执行短文本

过滤方法。网络接口用于与外部进行通信。本领域技术人员可以理解，图 7 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。

5 在一个实施例中，本申请提供的短文本过滤方法可以实现为一种计算机程序的形式，计算机程序可在如图 7 所示的计算机设备上运行。计算机设备的存储器中可存储组成短文本过滤装置的各个程序模板。比如，待检测文本获取模块 102、分词处理模块 104、文本过滤模块。

10 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，所述计算机程序被所述处理器执行时，使得所述处理器执行如下步骤：

 上述计算机设备，在其中一个实施例中，上述计算机程序被所述处理器执行时，还用于执行以下步骤：

 获取待检测文本，所述待检测文本为短文本；

 对所述待检测文本进行分词，获取包含至少一个词向量的分词结果；

15 计算所述分词结果与预设的语料数据库的相关度，根据所述相关度对所述待检测文本进行过滤。

 在其中一个实施例中，所述预设的语料数据库包含多条语料文本；上述计算机程序被所述处理器执行时，还用于执行以下步骤：按照预设的相似度计算算法分别计算所述分词结果与每一条语料文本之间的相似度；获取所述相似度
20 计算结果中的最大值为目标相似度，作为所述分词结果与预设的语料数据库的相关度。

 在其中一个实施例中，上述计算机程序被所述处理器执行时，还用于执行以下步骤：按照预设的编辑距离计算公式分别计算所述分词结果与每一条语料文本之间的编辑距离；获取所述编辑距离计算结果中的最大值为目标编辑距离，
25 作为所述分词结果与预设的语料数据库的相关度。

 在其中一个实施例中，上述计算机程序被所述处理器执行时，还用于执行以下步骤：确定所述目标相似度大于或等于预设的第一阈值；在所述目标相似度大于或等于所述第一阈值的情况下，判定所述待检测文本为相关文本；在所述目标相似度小于所述第一阈值的情况下，执行所述按照预设的编辑距离计算

公式分别计算所述分词结果与每一条语料文本之间的编辑距离；确定所述目标编辑距离大于或等于预设的第二阈值；判定所述待检测文本为相关文本。

5 在其中一个实施例中，上述计算机程序被所述处理器执行时，还用于执行以下步骤：在所述目标编辑距离小于所述第二阈值的情况下，确定所述目标相似度大于预设第三阈值、且目标编辑距离大于预设的第四阈值；在所述预设的关键词数据库查找与所述待检测文本匹配的关键词关键词，在查找到的情况下，判定所述待检测文本为相关文本。

10 在其中一个实施例中，所述对所述待检测文本进行分词的步骤之前，上述计算机程序被所述处理器执行时，还用于执行以下步骤：按照预设的文本改写算法对所述待检测文本进行改写，所述改写后的待检测文本中不包含预设的辅助词。

15 在其中一个实施例中，上述计算机程序被所述处理器执行时，还用于执行以下步骤：根据预设的关键词数据库对所述待检测文本进行分词，所述分词结果包括与所述预设的关键词数据库中包含的关键词对应的词向量以及多个字向量。

一种计算机可读存储介质，存储有计算机程序，所述计算机程序被处理器执行时，使得所述处理器执行如下步骤：

获取待检测文本，所述待检测文本为短文本；

对所述待检测文本进行分词，获取包含至少一个词向量的分词结果；

20 计算所述分词结果与预设的语料数据库的相关度，根据所述相关度对所述待检测文本进行过滤。

25 在其中一个实施例中，所述预设的语料数据库包含多条语料文本；上述计算机程序被所述处理器执行时，还用于执行以下步骤：按照预设的相似度计算算法分别计算所述分词结果与每一条语料文本之间的相似度；获取所述相似度计算结果中的最大值为目标相似度，作为所述分词结果与预设的语料数据库的相关度。

在其中一个实施例中，上述计算机程序被所述处理器执行时，还用于执行以下步骤：按照预设的编辑距离计算公式分别计算所述分词结果与每一条语料文本之间的编辑距离；获取所述编辑距离计算结果中的最大值为目标编辑距离，

作为所述分词结果与预设的语料数据库的相关度。

在其中一个实施例中，上述计算机程序被所述处理器执行时，还用于执行以下步骤：确定所述目标相似度大于或等于预设的第一阈值；

5 在所述目标相似度大于或等于所述第一阈值的情况下，判定所述待检测文本为相关文本；在所述目标相似度小于所述第一阈值的情况下，执行所述按照预设的编辑距离计算公式分别计算所述分词结果与每一条语料文本之间的编辑距离；确定所述目标编辑距离大于或等于预设的第二阈值；判定所述待检测文本为相关文本。

10 在其中一个实施例中，上述计算机程序被所述处理器执行时，还用于执行以下步骤：在所述目标编辑距离小于所述第二阈值的情况下，确定所述目标相似度大于预设第三阈值、且目标编辑距离大于预设的第四阈值；

在所述预设的关键词数据库查找与所述待检测文本匹配的关键词关键词，在查找到的情况下，判定所述待检测文本为相关文本。

15 在其中一个实施例中，上述计算机程序被所述处理器执行时，还用于执行以下步骤：按照预设的文本改写算法对所述待检测文本进行改写，所述改写后的待检测文本中不包含预设的辅助词。

20 在其中一个实施例中，上述计算机程序被所述处理器执行时，还用于执行以下步骤：根据预设的关键词数据库对所述待检测文本进行分词，所述分词结果包括与所述预设的关键词数据库中包含的关键词对应的词向量以及多个字向量。

需要说明的是，上述短文本过滤方法、短文本过滤装置、计算机设备和计算机可读存储介质属于同一个发明构思，短文本过滤方法、短文本过滤装置、计算机设备和计算机可读存储介质中涉及的内容可相互适用。

25 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，是可以通过计算机程序来指令相关的硬件来完成，所述的程序可存储于一非易失性计算机可读取存储介质中，该程序在执行时，可包括如上述各方法的实施例的流程。其中，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本申请专利范围的限制。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

1、一种短文本过滤方法，其特征在于，所述方法包括：

获取待检测文本，所述待检测文本为短文本；

对所述待检测文本进行分词，获取包含至少一个词向量的分词结果；

5 计算所述分词结果与预设的语料数据库的相关度，根据所述相关度对所述待检测文本进行过滤。

2、根据权利要求 1 所述的短文本过滤方法，其特征在于，所述预设的语料数据库包含多条语料文本；

所述计算所述分词结果与预设的语料数据库的相关度，还包括：

10 按照预设的相似度计算算法分别计算所述分词结果与每一条语料文本之间的相似度；

获取所述相似度计算结果中的最大值为目标相似度，作为所述分词结果与预设的语料数据库的相关度。

15 3、根据权利要求 2 所述的短文本过滤方法，其特征在于，所述计算所述分词结果与预设的语料数据库的相关度，还包括：

按照预设的编辑距离计算公式分别计算所述分词结果与每一条语料文本之间的编辑距离；

获取所述编辑距离计算结果中的最大值为目标编辑距离，作为所述分词结果与预设的语料数据库的相关度。

20 4、根据权利要求 3 所述的短文本过滤方法，其特征在于，所述根据所述相关度对所述待检测文本进行过滤，还包括：

确定所述目标相似度大于或等于预设的第一阈值；

在所述目标相似度大于或等于所述第一阈值的情况下，判定所述待检测文本为相关文本；

25 在所述目标相似度小于所述第一阈值的情况下，执行所述按照预设的编辑距离计算公式分别计算所述分词结果与每一条语料文本之间的编辑距离；确定所述目标编辑距离大于或等于预设的第二阈值；判定所述待检测文本为相关文本。

5、根据权利要求4所述的短文本过滤方法，其特征在于，所述确定所述目标编辑距离大于或等于预设的第二阈值之后，还包括：

在所述目标编辑距离小于所述第二阈值的情况下，确定所述目标相似度大于预设第三阈值、且目标编辑距离大于预设的第四阈值；

5 在所述预设的关键词数据库查找与所述待检测文本匹配的关键词关键词，在查找到的情况下，判定所述待检测文本为相关文本。

6、根据权利要求1所述的短文本过滤方法，其特征在于，所述对所述待检测文本进行分词的步骤之前还包括：

按照预设的文本改写算法对所述待检测文本进行改写，所述改写后的待检测文本中不包含预设的辅助词。

7、根据权利要求1所述的短文本过滤方法，其特征在于，所述对所述待检测文本进行分词，还包括：

根据预设的关键词数据库对所述待检测文本进行分词，所述分词结果包括与所述预设的关键词数据库中包含的关键词对应的词向量以及多个字向量。

15 8、一种短文本过滤装置，其特征在于，所述装置包括：

待检测文本获取模块，用于获取待检测文本，所述待检测文本为短文本；
分词处理模块，用于对所述待检测文本进行分词，获取包含至少一个词向量的分词结果；

文本过滤模块，用于计算所述分词结果与预设的语料数据库的相关度，根据所述相关度对所述待检测文本进行过滤。

9、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，所述计算机程序被所述处理器执行时，使得所述处理器执行如权利要求1至7中任一项所述方法的步骤。

10、一种计算机可读存储介质，存储有计算机程序，所述计算机程序被处理器执行时，使得所述处理器执行如权利要求1至7中任一项所述方法的步骤。

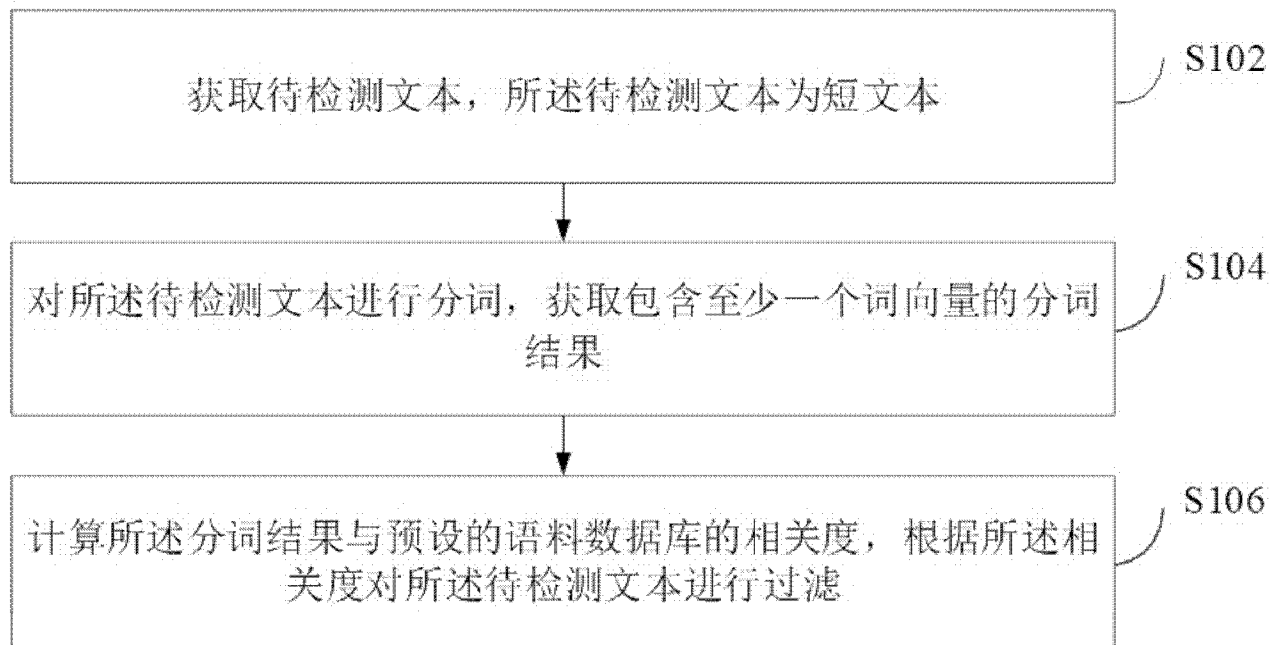


图 1

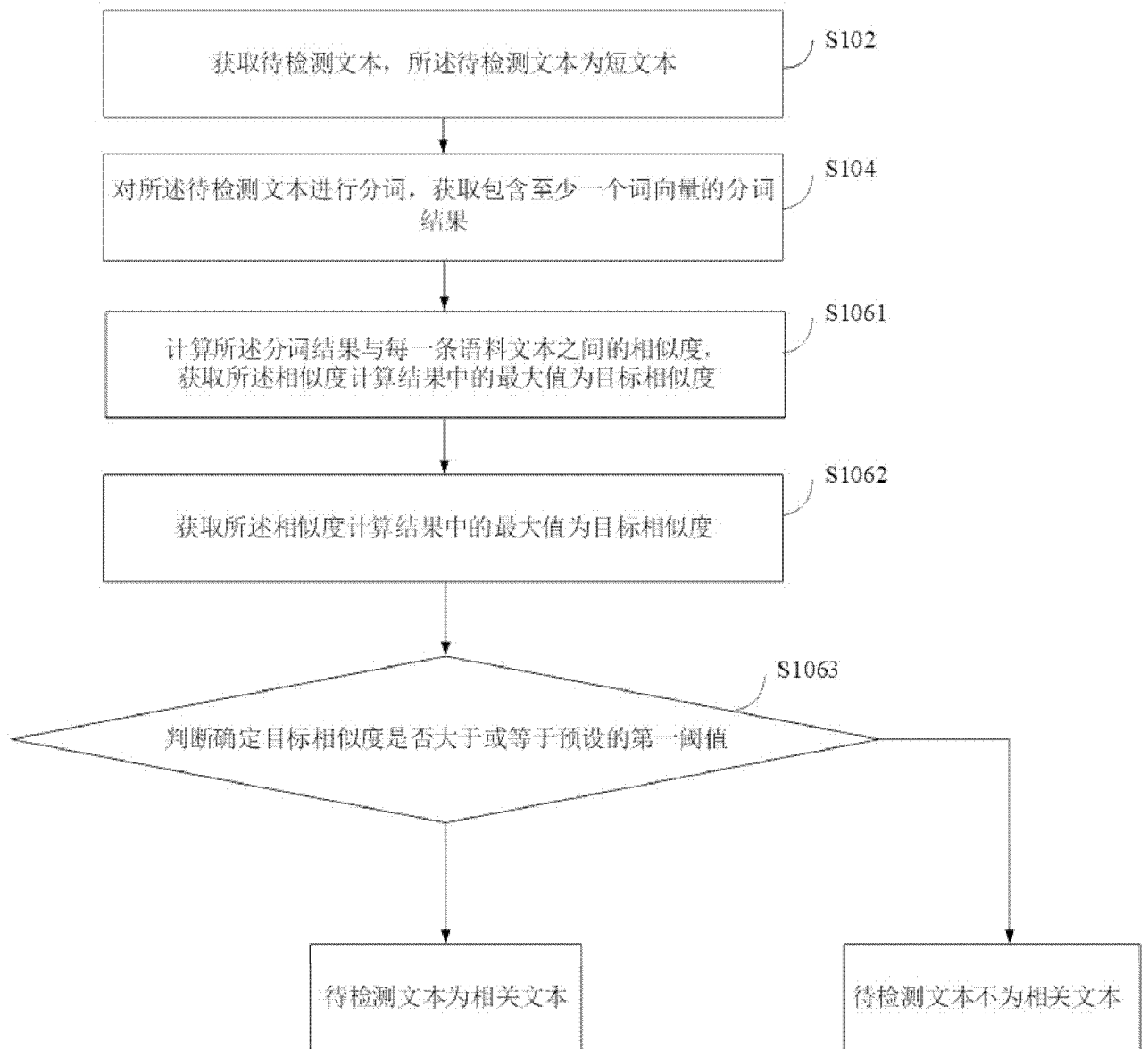


图 2

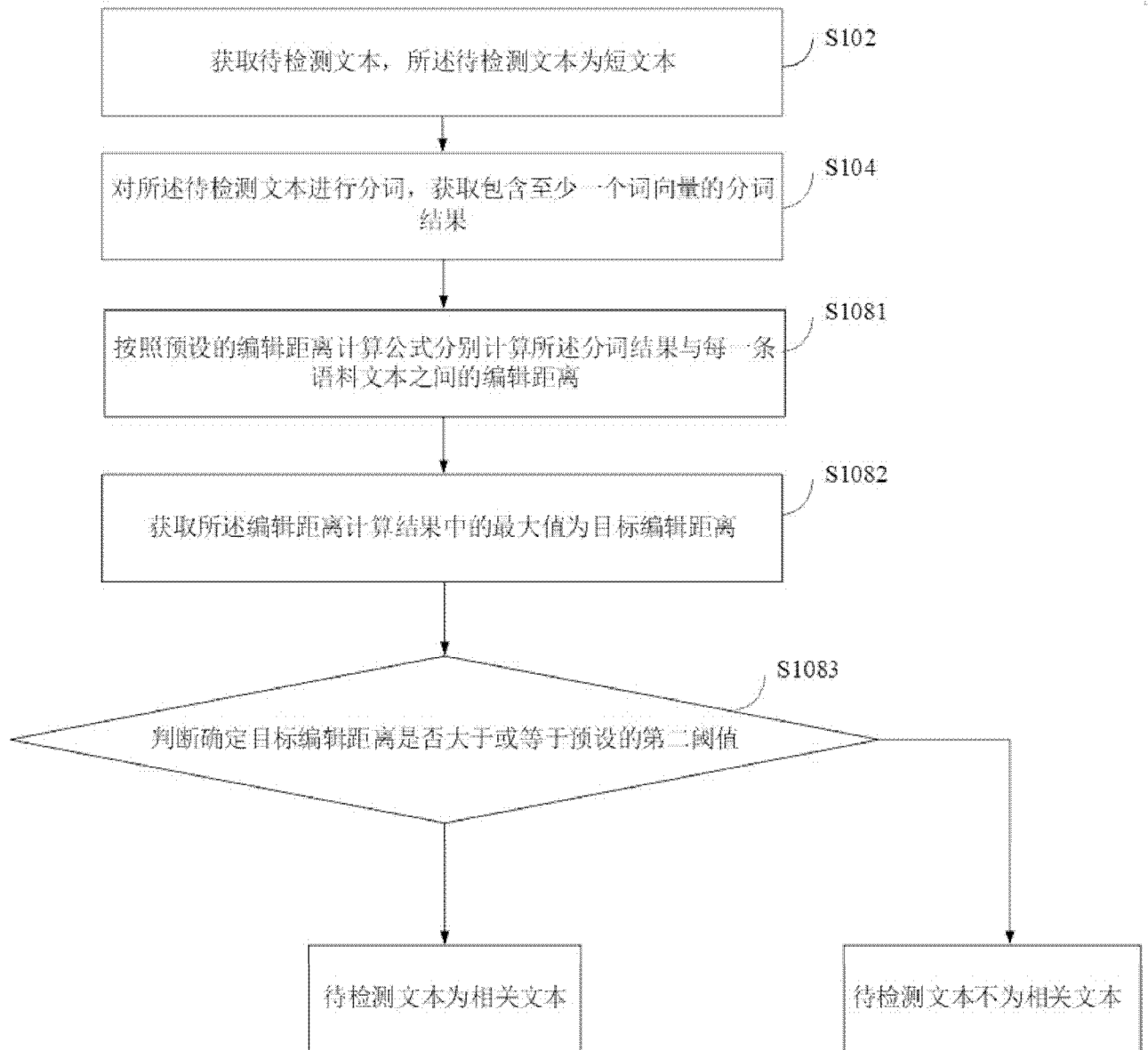


图 3

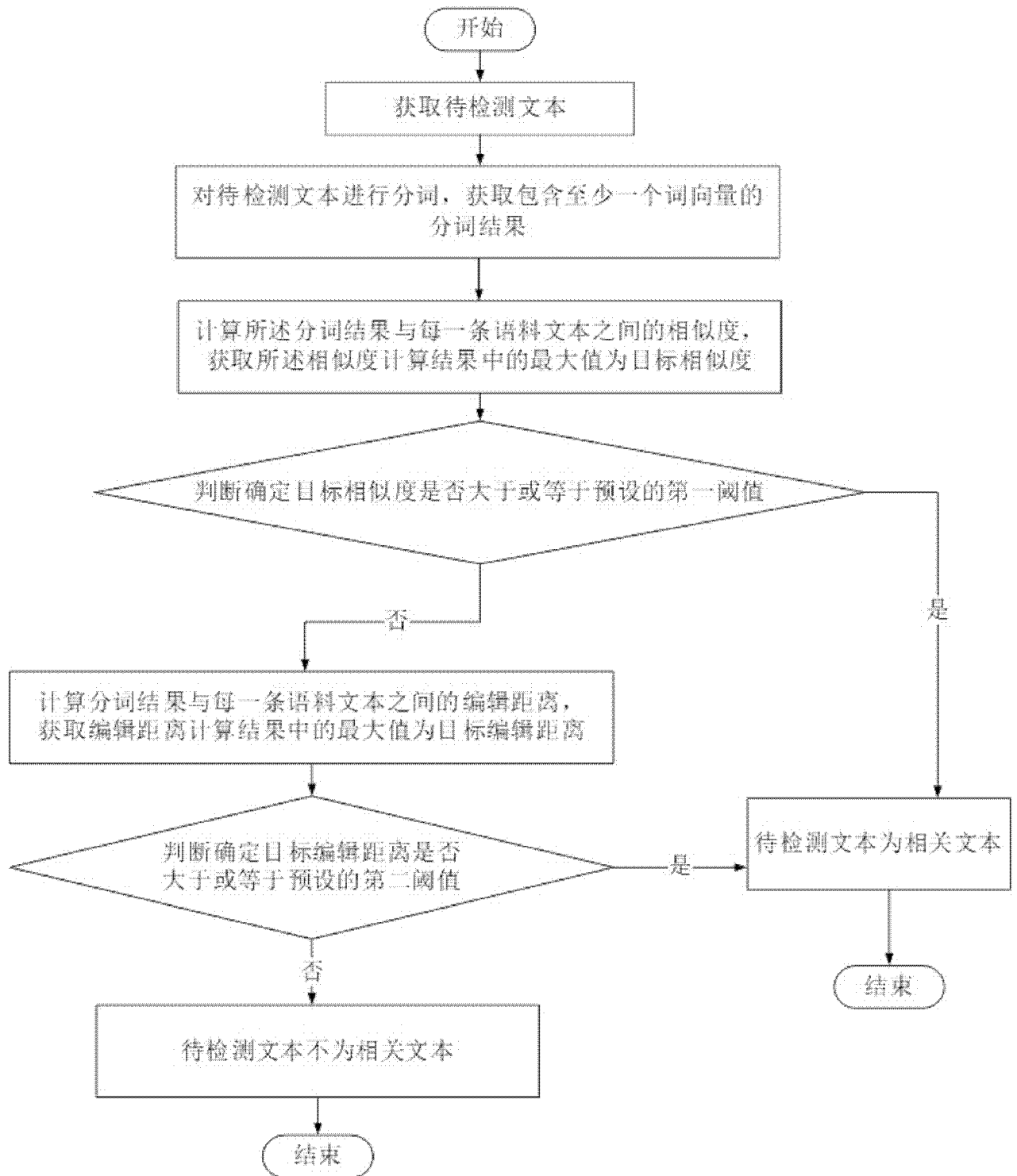


图 4

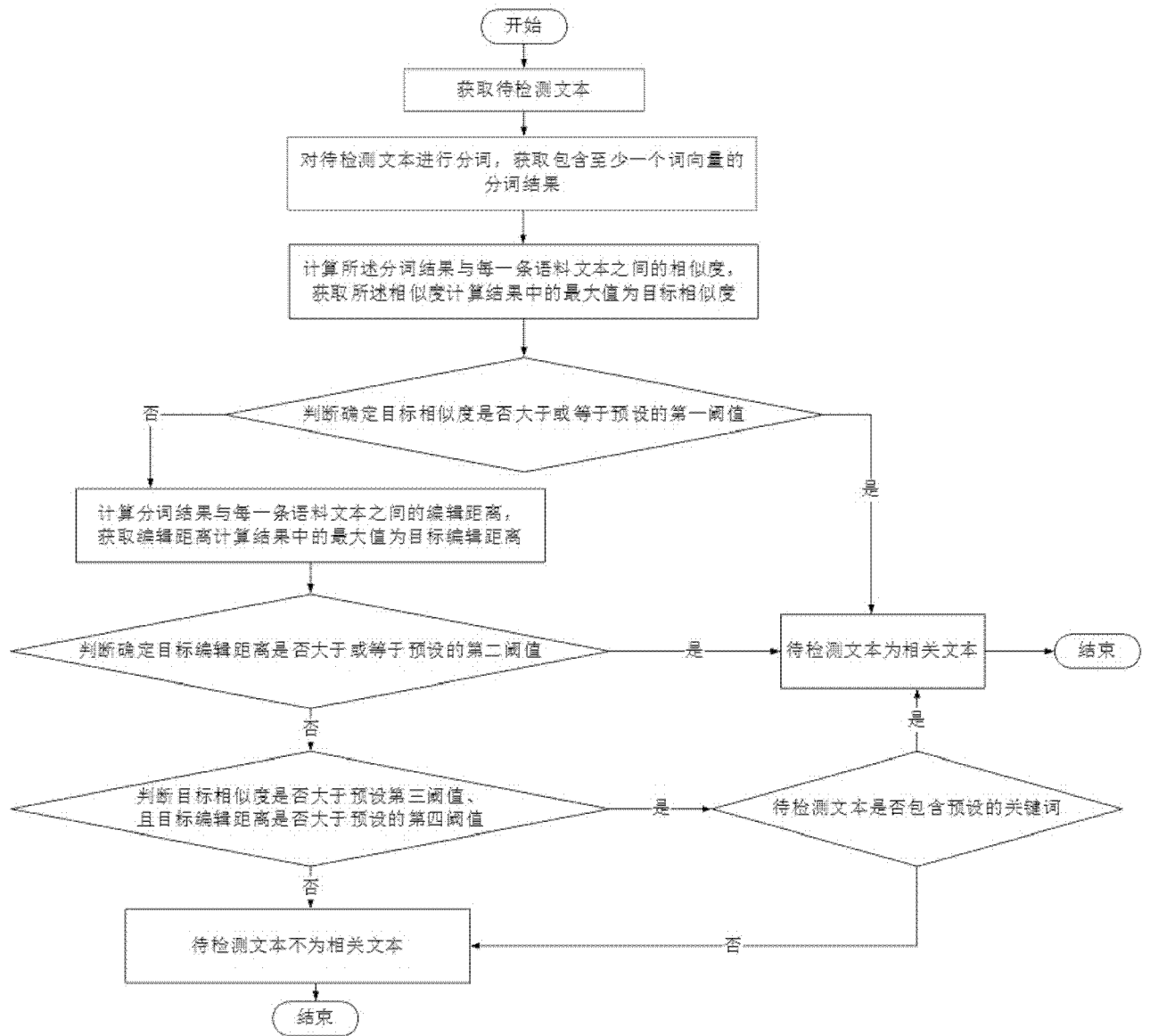


图 5

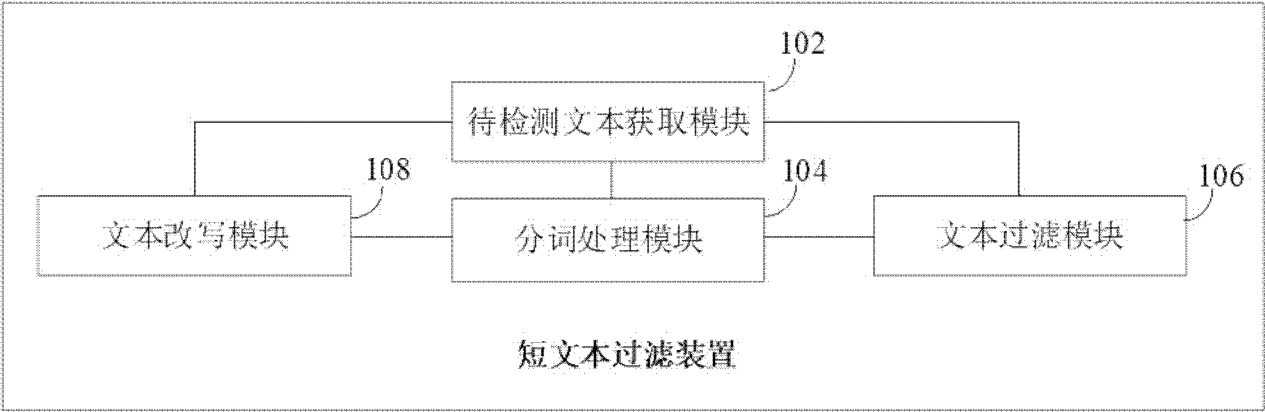


图 6

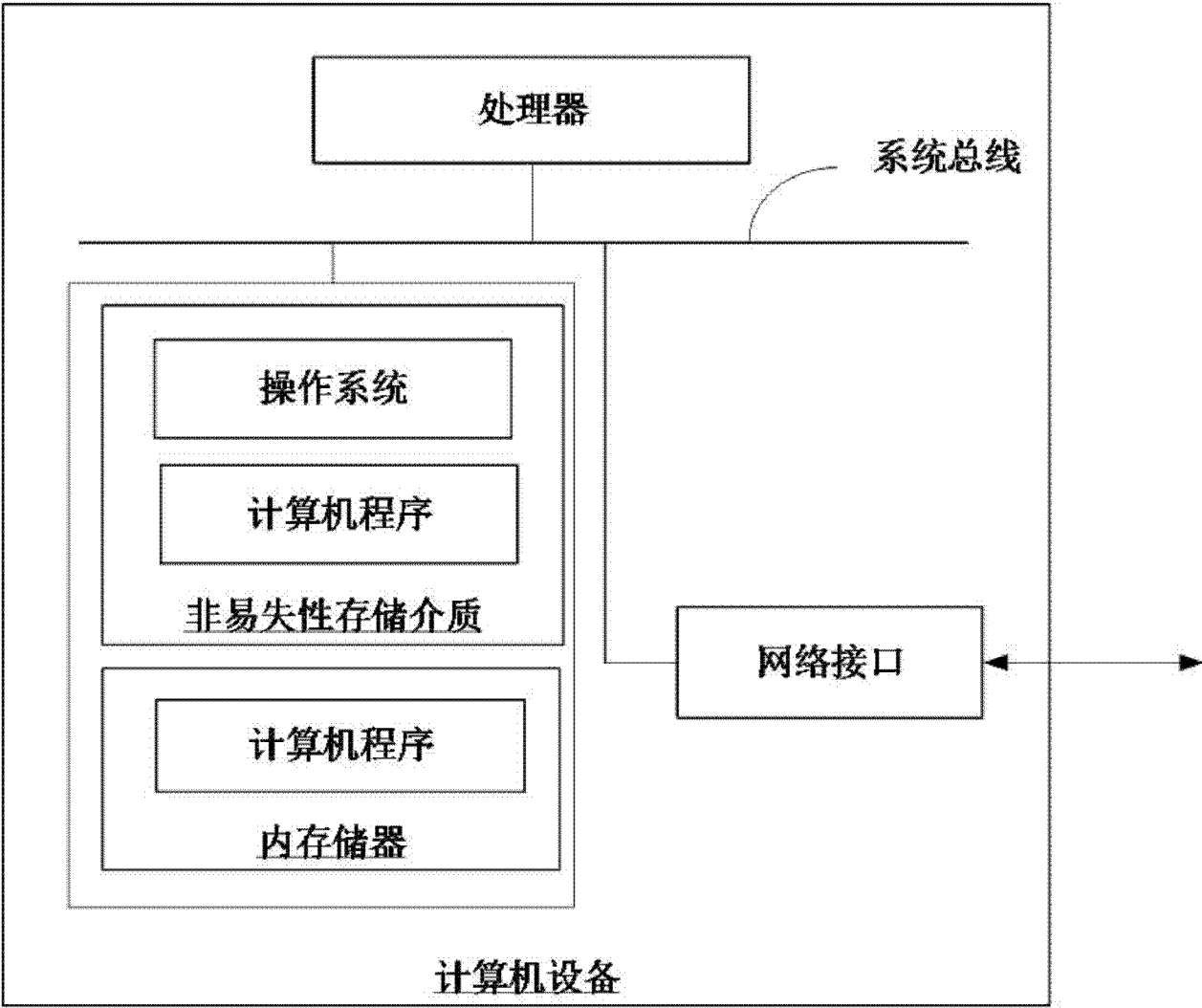


图 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/123847

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/335(2019.01)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)

CNPAT, WPI, EPODOC, IEEE, CNKI: 短文本, 文本, 过滤, 分类, 分词, 相似, 相关度, 编辑距离, 阈值, 辅助词, short text, text, filter, classification, word, segmentation, similarity, relevance, edit distance, threshold, auxiliary word

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 108763477 A (XIAMEN KUAISHANGTONG INFORMATION TECHNOLOGY CO., LTD.) 06 November 2018 (2018-11-06) description, paragraphs [0035]-[0041], [0062] and [0076]-[0082]	1-10
A	CN 104102626 A (XIAMEN TUIITE INFORMATION TECHNOLOGY CO., LTD.) 15 October 2014 (2014-10-15) entire document	1-10
A	CN 107085581 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 22 August 2017 (2017-08-22) entire document	1-10
A	CN 101621391 A (BEIJING BAIWEN BAIDA NETWORK TECHNOLOGIES CO., LTD.) 06 January 2010 (2010-01-06) entire document	1-10
A	US 2009240498 A1 (MICROSOFT CORPORATION) 24 September 2009 (2009-09-24) entire document	1-10



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

04 June 2019

Date of mailing of the international search report

28 June 2019

Name and mailing address of the ISA/CN

State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Authorized officer

Facsimile No. (86-10)62019451

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/123847

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	108763477	A	06 November 2018	None	
CN	104102626	A	15 October 2014	CN 104102626	B 15 August 2017
CN	107085581	A	22 August 2017	None	
CN	101621391	A	06 January 2010	None	
US	2009240498	A1	24 September 2009	None	

国际检索报告

国际申请号

PCT/CN2018/123847

A. 主题的分类

G06F 16/335(2019.01)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))

CNPAT, WPI, EPDOC, IEEE, CNKI: 短文本, 文本, 过滤, 分类, 分词, 相似, 相关度, 编辑距离, 阈值, 辅助词, short text, text, filter, classification, word, segmentation, similarity, relevance, edit distance, threshold, auxiliary word

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
X	CN 108763477 A (厦门快商通信息技术有限公司) 2018年 11月 6日 (2018 - 11 - 06) 说明书第[0035]-[0041], [0062], [0076]-[0082]段	1-10
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) 全文	1-10
A	CN 107085581 A (腾讯科技深圳有限公司) 2017年 8月 22日 (2017 - 08 - 22) 全文	1-10
A	CN 101621391 A (北京百问百答网络技术有限公司) 2010年 1月 6日 (2010 - 01 - 06) 全文	1-10
A	US 2009240498 A1 (MICROSOFT CORPORATION) 2009年 9月 24日 (2009 - 09 - 24) 全文	1-10

☐ 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 6月 4日

国际检索报告邮寄日期

2019年 6月 28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

石爽

传真号 (86-10)62019451

电话号码 86-(10)-53961413

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/123847

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	108763477	A	2018年 11月 6日	无			
CN	104102626	A	2014年 10月 15日	CN	104102626	B	2017年 8月 15日
CN	107085581	A	2017年 8月 22日	无			
CN	101621391	A	2010年 1月 6日	无			
US	2009240498	A1	2009年 9月 24日	无			