(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112364620 A
(43)申请公布日 2021.02.12

(21)申请号 202011231992.0

(22)申请日 2020.11.06

(71)申请人 中国平安人寿保险股份有限公司
地址 518000 广东省深圳市福田区益田路5033号平安金融中心14、15、16、41、44、45、46层

(72)发明人 杨威

(74)专利代理机构 深圳市明日今典知识产权代理事务所(普通合伙) 44343
代理人 王杰辉 曹勇

(51)Int.Cl.
G06F 40/194(2020.01)
G06F 40/284(2020.01)
G06N 20/10(2019.01)

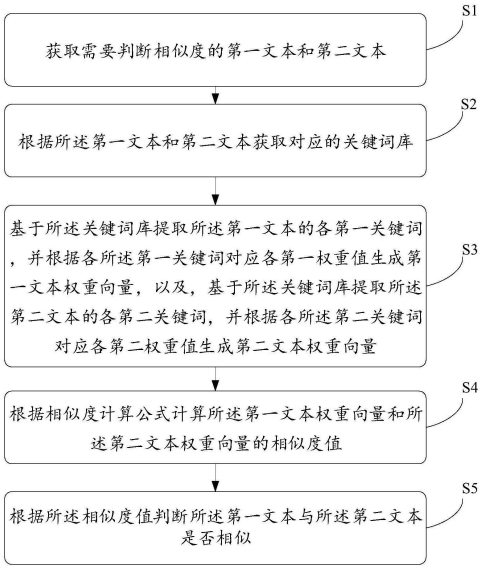
权利要求书3页 说明书9页 附图3页

(54)发明名称

文本相似度的判断方法、装置以及计算机设备

(57)摘要

本发明提供了一种文本相似度的判断方法、装置以及计算机设备，其中，方法包括：获取需要判断相似度的第一文本和第二文本；根据所述第一文本和第二文本获取对应的关键词库；基于所述关键词库提取文本的各关键词，并根据各关键词对应各权重值生成文本权重向量；根据相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值；根据所述相似度值判断所述第一文本与所述第二文本是否相似。本发明的有益效果：通过为不同的关键词设置不同的权重，并且用权重来将第一文本和第二文本向量化，使其基于关键词的权重值计算相似度，从而可以提高第一文本和第二文本相似度判断与人为判断更接近。



CN 112364620 A

1.一种文本相似度的判断方法，其特征在于，包括：

获取需要判断相似度的第一文本和第二文本；

根据所述第一文本和第二文本获取对应的关键词库，其中，所述关键词库中存储有多个关键词，以及与各所述关键词一一对应的权重值；

基于所述关键词库提取所述第一文本的各第一关键词，并根据各所述第一关键词对应各第一权重值生成第一文本权重向量，以及，基于所述关键词库提取所述第二文本的各第二关键词，并根据各所述第二关键词对应各第二权重值生成第二文本权重向量；

根据相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值；

根据所述相似度值判断所述第一文本与所述第二文本是否相似。

2.如权利要求1所述的文本相似度的判断方法，其特征在于，所述根据相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值的步骤，包括：

分别获取所述第一文本权重向量和所述第二文本权重向量中各关键词的个数；

根据各关键词对应的所述权重以及个数，通过所述相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值；其中，所述相似度计算公式为 $cos(I,R)=\frac{\sum_{i=1}^{n} f(x_i)*f(y_i)}{\sqrt{\sum_{i=1}^{n} f^2(x_i)}*\sqrt{\sum_{i=1}^{n} f^2(y_i)}}$，I表示所述第一文本权重向量，R表示第二文本权重向量，cos(I,R)表示所述相似度值，x_i表示所述第一文本权重向量的第i个关键词对应的个数，y_i表示所述第二文本权重向量的第i个关键词对应的个数，n表示所述关键词库中关键词的个数，$f(x_i)=\begin{cases}0, x=0\\ w_i, x>0\end{cases}$，$f(y_i)=\begin{cases}0, x=0\\ w_i, x>0\end{cases}$，$w_i$表示第个关键词对应的权重。

3.如权利要求1所述的文本相似度的判断方法，其特征在于，所述根据所述第一文本和第二文本获取对应的关键词库的步骤之前，还包括：

按照预设的规则将所述关键词划分为多个类别；

将多个类别中的第一类关键词取出，并为其他关键词按照预设的权重规则划分权重；

将多组相似文本依次输入至权重训练模型中进行训练，得到所述第一类关键词的权重参数。

4.如权利要求1所述的文本相似度的判断方法，其特征在于，所述基于所述关键词库提取所述第一文本的各第一关键词，并根据各所述第一关键词对应各第一权重值生成第一文本权重向量，以及，基于所述关键词库提取所述第二文本的各第二关键词，并根据各所述第二关键词对应各第二权重值生成第二文本权重向量的步骤，包括：

通过文本分类器分别对所述第一文本和所述第二文本进行分词，分别对应得到第一词语和第二词语；

根据所述关键词库将所述第一词语中对应的所述第一关键词，以及所述第二词语中对应的所述第二关键词提取出来；

根据所述第一关键词对应各第一权重值生成第一文本权重向量，以及根据所述第二关键词对应各第二权重值生成第二文本权重向量。

5.如权利要求1所述的文本相似度的判断方法，其特征在于，所述根据所述第一文本和第二文本获取对应的关键词库的步骤，包括：

将所述第一文本和所述第二文本依次输入至自然语言处理后的机器学习模型中，并计算得到所述第一文本与所述第二文本与各个类别分别对应的第一类别相似度和第二类别相似度；

提取所述第一类别相似度中大于类别预设相似度的第一期望类别，以及提取所述第二类别相似度中大于类别预设相似度的第二期望类别；

提取所述第一期望类别中与所述第二期望类别中相同的目标类别，并获取所述目标类别对应的关键词库。

6.如权利要求1所述的文本相似度的判断方法，其特征在于，所述根据所述第一文本和第二文本获取对应的关键词库的步骤之前，还包括：

按照预设的规则将所述关键词划分为多个等级；

设置最低等级的关键词的最低权重，以及根据公式$w_k = R_t + \sqrt{\sum_{t=k+1}^{c} n_t w_t^2}$设定其余等级的所述关键词的权重，其中，$w_c$表示所述最低权重，$w_t$表示第t等级的权重，$R_t$表示第t等级的预设参数，$n_t$表示第t等级所有关键词的总数量。

7.一种文本相似度的判断装置，其特征在于，包括：

文本获取模块，用于获取需要判断相似度的第一文本和第二文本；

关键词库获取模块，用于根据所述第一文本和第二文本获取对应的关键词库，其中，所述关键词库中存储有多个关键词，以及与各所述关键词一一对应的权重值；

权重向量生成模块，用于基于所述关键词库提取所述第一文本的各第一关键词，并根据各所述第一关键词对应各第一权重值生成第一文本权重向量，以及，基于所述关键词库提取所述第二文本的各第二关键词，并根据各所述第二关键词对应各第二权重值生成第二文本权重向量；

相似度值计算模块，用于根据相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值；

相似判断模块，用于根据所述相似度值判断所述第一文本与所述第二文本是否相似。

8.如权利要求7所述的文本相似度的判断装置，其特征在于，所述相似度值计算模块，包括：

个数获取子模块，用于分别获取所述第一文本权重向量和所述第二文本权重向量中各关键词的个数；

相似度值计算子模块，用于根据各关键词对应的所述权重以及个数，通过所述相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值；其中，所述相似度计算公式为$cos(I,R) = \frac{\sum_{i=1}^{n} f(x_i) * f(y_i)}{\sqrt{\sum_{i=1}^{n} f^2(x_i)} * \sqrt{\sum_{i=1}^{n} f^2(y_i)}}$，I表示所述第一文本权重向量，R表示第二文本权重向量，cos(I,R)表示所述相似度值，x_i表示所述第一文本权重

向量的第i个关键词对应的个数，y_i表示所述第二文本权重向量的第i个关键词对应的个数，n表示所述关键词库中关键词的个数，$f\left(x_i\right)=\begin{cases}0, x=0\\ w_i, x>0\end{cases}$，$f\left(y_i\right)=\begin{cases}0, x=0\\ w_i, x>0\end{cases}$，$w_i$表示第个关键词对应的权重。

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，其特征在于，所述处理器执行所述计算机程序时实现权利要求1至6中任一项所述方法的步骤。

10.一种计算机可读存储介质，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权利要求1至6中任一项所述的方法的步骤。

文本相似度的判断方法、装置以及计算机设备

技术领域

[0001] 本发明涉及语言处理领域，特别涉及一种文本相似度的判断方法、装置以及计算机设备。

背景技术

[0002] 目前检测两个文本的相似度都是根据预设的向量机将两个文本进行向量化，通过余弦相似度算法计算两个向量的相似度，但是随着文本的长度不同，即使是相关的两个文本也会计算为不相关，现有技术中，一般都是将文本划分为多个向量，然后检测其中相同的向量数目，但是在文本表述中，尤其是由同一个人的表述，其撰写两个文本的风格相近，即使是完全不同的类别也会有很高的相似度，文本相似度计算不准确，因此亟需一种文本相似度的判断方法。

发明内容

[0003] 本发明的主要目的为提供一种文本相似度的判断方法、装置以及计算机设备，旨在解决文本相似度计算不准确的问题。

[0004] 本发明提供了一种文本相似度的判断方法，包括：

[0005] 获取需要判断相似度的第一文本和第二文本；

[0006] 根据所述第一文本和第二文本获取对应的关键词库，其中，所述关键词库中存储有多个关键词，以及与各所述关键词一一对应的权重值；

[0007] 基于所述关键词库提取所述第一文本的各第一关键词，并根据各所述第一关键词对应各第一权重值生成第一文本权重向量，以及，基于所述关键词库提取所述第二文本的各第二关键词，并根据各所述第二关键词对应各第二权重值生成第二文本权重向量；

[0008] 根据相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值；

[0009] 根据所述相似度值判断所述第一文本与所述第二文本是否相似。

[0010] 进一步地，所述根据相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值的步骤，包括：

[0011] 分别获取所述第一文本权重向量和所述第二文本权重向量中各关键词的个数；

[0012] 根据各关键词对应的所述权重以及个数，通过所述相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值；其中，所述相似度计算公式为 $cos(I,R)=\frac{\sum_{i=1}^{n} f(x_i)*f(y_i)}{\sqrt{\sum_{i=1}^{n} f^2(x_i)}*\sqrt{\sum_{i=1}^{n} f^2(y_i)}}$，I表示所述第一文本权重向量，R表示第二文本权重向量，cos(I,R)表示所述相似度值，x_i表示所述第一文本权重向量的第i个关键词对应的个数，y_i表示所述第二文本权重向量的第i个关键词对应的个数，n表示所述关键

词库中关键词的个数，$f\left(x_i\right)=\begin{cases}0, x=0\\ w_i, x>0\end{cases}$，$f\left(y_i\right)=\begin{cases}0, x=0\\ w_i, x>0\end{cases}$，$w_i$表示第个关键词对应的权重。

[0013] 进一步地，所述根据所述第一文本和第二文本获取对应的关键词库的步骤之前，还包括：

[0014] 按照预设的规则将所述关键词划分为多个类别；

[0015] 将多个类别中的第一类关键词取出，并为其他关键词按照预设的权重规则划分权重；

[0016] 将多组相似文本依次输入至权重训练模型中进行训练，得到所述第一类关键词的权重参数。

[0017] 进一步地，所述基于所述关键词库提取所述第一文本的各第一关键词，并根据各所述第一关键词对应各第一权重值生成第一文本权重向量，以及，基于所述关键词库提取所述第二文本的各第二关键词，并根据各所述第二关键词对应各第二权重值生成第二文本权重向量的步骤，包括：

[0018] 通过文本分类器分别对所述第一文本和所述第二文本进行分词，分别对应得到第一词语和第二词语；

[0019] 根据所述关键词库将所述第一词语中对应的所述第一关键词，以及所述第二词语中对应的所述第二关键词提取出来；

[0020] 根据所述第一关键词对应各第一权重值生成第一文本权重向量，以及根据所述第二关键词对应各第二权重值生成第二文本权重向量。

[0021] 进一步地，所述根据所述第一文本和第二文本获取对应的关键词库的步骤，包括：

[0022] 将所述第一文本和所述第二文本依次输入至自然语言处理后的机器学习模型中，并计算得到所述第一文本与所述第二文本与各个类别分别对应的第一类别相似度和第二类别相似度；

[0023] 提取所述第一类别相似度中大于类别预设相似度的第一期望类别，以及提取所述第二类别相似度中大于类别预设相似度的第二期望类别；

[0024] 提取所述第一期望类别中与所述第二期望类别中相同的目标类别，并获取所述目标类别对应的关键词库。

[0025] 进一步地，所述根据所述第一文本和第二文本获取对应的关键词库的步骤之前，还包括：

[0026] 按照预设的规则将所述关键词划分为多个等级；

[0027] 设置最低等级的关键词的最低权重，以及根据公式$W_k=R_t+\sqrt{\sum_{t=k+1}^{c}n_t{w_t}^2}$设定其余等级的所述关键词的权重，其中，$w_c$表示所述最低权重，$w_t$表示第t等级的权重，$R_t$表示第t等级的预设参数，$n_t$表示第t等级所有关键词的总数量。

[0028] 本发明提供了一种文本相似度的判断装置，包括：

[0029] 文本获取模块，用于获取需要判断相似度的第一文本和第二文本；

[0030] 关键词库获取模块，用于根据所述第一文本和第二文本获取对应的关键词库，其中，所述关键词库中存储有多个关键词，以及与各所述关键词一一对应的权重值；

[0031] 权重向量生成模块，用于基于所述关键词库提取所述第一文本的各第一关键词，并根据各所述第一关键词对应各第一权重值生成第一文本权重向量，以及，基于所述关键词库提取所述第二文本的各第二关键词，并根据各所述第二关键词对应各第二权重值生成第二文本权重向量；

[0032] 相似度值计算模块，用于根据相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值；

[0033] 相似判断模块，用于根据所述相似度值判断所述第一文本与所述第二文本是否相似。

[0034] 进一步地，所述相似度值计算模块，包括：

[0035] 个数获取子模块，用于分别获取所述第一文本权重向量和所述第二文本权重向量中各关键词的个数；

[0036] 相似度值计算子模块，用于根据各关键词对应的所述权重以及个数，通过所述相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值；其中，所述相似度计算公式为$cos(I,R)=\frac{\sum_{i=1}^{n} f(x_i)*f(y_i)}{\sqrt{\sum_{i=1}^{n} f^2(x_i)}*\sqrt{\sum_{i=1}^{n} f^2(y_i)}}$，I表示所述第一文本权重向量，R表示第二文本权重向量，cos(I,R)表示所述相似度值，x_i表示所述第一文本权重向量的第i个关键词对应的个数，y_i表示所述第二文本权重向量的第i个关键词对应的个数，n表示所述关键词库中关键词的个数，$f(x_i)=\begin{cases}0, x=0\\ w_i, x>0\end{cases}$，$f(y_i)=\begin{cases}0, x=0\\ w_i, x>0\end{cases}$，$w_i$表示第个关键词对应的权重。

[0037] 本发明提供了一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述处理器执行所述计算机程序时实现上述任一项所述方法的步骤。

[0038] 本发明还提供了一种计算机可读存储介质，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。

[0039] 本发明的有益效果：通过为不同的关键词设置不同的权重，并且用权重来将第一文本和第二文本向量化，使其基于关键词的权重值计算相似度，从而可以提高第一文本和第二文本相似度判断与人为判断更接近。

附图说明

[0040] 图1是本发明一实施例的一种文本相似度的判断方法的流程示意图；

[0041] 图2是本发明一实施例的一种文本相似度的判断装置的结构示意框图；

[0042] 图3为本申请一实施例的计算机设备的结构示意框图。

[0043] 本发明目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

[0044] 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的一部分实施例，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。

[0045] 需要说明，本发明实施例中所有方向性指示(诸如上、下、左、右、前、后等)仅用于解释在某一特定姿态(如附图所示)下各部件之间的相对位置关系、运动情况等，如果该特定姿态发生改变时，则该方向性指示也相应地随之改变，所述的连接可以是直接连接，也可以是间接连接。

[0046] 本文中术语“和/或”，仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A和B，可以表示：单独存在A，同时存在A和B，单独存在B这三种情况。

[0047] 另外，在本发明中如涉及“第一”、“第二”等的描述仅用于描述目的，而不能理解为指示或暗示其相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。另外，各个实施例之间的技术方案可以相互结合，但是必须是以本领域普通技术人员能够实现为基础，当技术方案的结合出现相互矛盾或无法实现时应当认为这种技术方案的结合不存在，也不在本发明要求的保护范围之内。

[0048] 参照图1，本发明提出一种文本相似度的判断方法，包括：

[0049] S1：获取需要判断相似度的第一文本和第二文本；

[0050] S2：根据所述第一文本和第二文本获取对应的关键词库，其中，所述关键词库中存储有多个关键词，以及与各所述关键词一一对应的权重值；

[0051] S3：基于所述关键词库提取所述第一文本的各第一关键词，并根据各所述第一关键词对应各第一权重值生成第一文本权重向量，以及，基于所述关键词库提取所述第二文本的各第二关键词，并根据各所述第二关键词对应各第二权重值生成第二文本权重向量；

[0052] S4：根据相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值；

[0053] S5：根据所述相似度值判断所述第一文本与所述第二文本是否相似。

[0054] 如上述步骤S1所述，获取需要判断相似度的第一文本和第二文本。具体地，第一文本和第二文本可以是用户上传的文本，也可以是从其他APP上下载的文本，常见的文本文档的扩展名有txt、doc、docx、wps。

[0055] 如上述步骤S2所述，根据所述第一文本和第二文本获取对应的关键词库。具体地，可以根据第一文本和第二文本的文本信息，找出与第一文本和第二文本都接近的关键词库，也可以是基于第一文本或第二文本找出对应最接近的关键词库。应当理解，关键词库是作为第一文本和第二文本相似度判断的主要依据，若关键词库中的关键词与第一文本以及第二文本中携带的文本信息的关联度都比较差，则后续判断相似度时的误差也会比较大，故而找出的关键词库应当至少与第一文本和第二文本中的一个文本相关。

[0056] 其中，关键词库中设置关键词权重的方法为，事先设置好关键词与权重的对应关系，然后根据对应关系为每个关键词设置对应的权重，需要说明的是，权重是一个具体的数值，若其中一个关键词对第一文本和第二文本之间的相似度值影响较大，则该关键词对应的权重数值也相应较大，若其中一个关键词对第一文本和第二文本之间的相似度值影响较小，则该关键词对应的权重数值也相应较小。需要说明的是，第一关键词或第二关键词是指

第一文本或者第二文本中所有的关键词，并不是指只有一个关键词。一般而言，第一关键词和第二关键词分别对应多个关键词。

[0057] 如上述步骤S3所述，根据各个关键词对应的权重，将第一文本和第二文本进行向量化处理，即可以先通过文本分类器将第一文本和第二文本分别进行分词处理，根据各个关键词对应的权重按照文本顺序将第一文本和第二文本进行向量化，分别得到对应的第一文本权重向量和第二文本权重向量。其中，第一文本权重向量和第二文本权重向量的各个维度由不同的关键词的权重构成，各权重的维度位置由关键词在文本中的位置顺序所决定。当然在一些计算相似度的实施例中，维度的前后顺序不影响其最终的相似度值，即将维度的位置相互调换也不影响最终的相似度值。故而也可以根据对各个权重对应的维度进行随机排列，得到对应的第一文本权重向量和第二文本权重向量。

[0058] 如上述步骤S4所述，根据相似度计算公式计算所述第一文本权重向量和第二文本权重向量之间的相似度值，应当理解的是，该相似度计算公式不宜使用每一维对比计算的公式，优选使用余弦相似度计算公式，其中，相似度值越高表明第一文本和第二文本越相似，相似度值越低表明第一文本和第二文本越不相似。

[0059] 如上述步骤S5所述，根据相似度值判断第一文本和第二文本是否相似，其中，判断的方式可以是事先设定相似度阈值，当相似度值大于或等于该阈值时，表明第一文本和第二文本相似，当相似度值小于该阈值时，表明第一文本和第二文本不相似。

[0060] 在一个实施例中，所述根据相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值的步骤S4，包括：

[0061] S401：分别获取所述第一文本权重向量和所述第二文本权重向量中各关键词的个数；

[0062] S402：根据各关键词对应的所述权重以及个数，通过所述相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值；其中，所述相似度计算公式为 $cos(I,R)=\frac{\sum_{i=1}^{n} f(x_i)*f(y_i)}{\sqrt{\sum_{i=1}^{n} f^2(x_i)}*\sqrt{\sum_{i=1}^{n} f^2(y_i)}}$，I表示所述第一文本权重向量，R表示第二文本权重向量，cos(I,R)表示所述相似度值，x_i表示所述第一文本权重向量的第i个关键词对应的个数，y_i表示所述第二文本权重向量的第i个关键词对应的个数，n表示所述关键词库中关键词的个数，$f(x_i)=\begin{cases}0, x=0\\ w_i, x>0\end{cases}$，$f(y_i)=\begin{cases}0, x=0\\ w_i, x>0\end{cases}$，$w_i$表示第个关键词对应的权重。

[0063] 如上述步骤S401-S402所述，由于关键词之间的顺序关系对于第一文本权重向量和第二文本权重向量之间的相似度判断影响较低，因此可以忽略掉关键词之间的顺序关系，根据各个关键词的个数以及对应的权重进行计算，相似度计算公式为 $cos(I,R)=\frac{\sum_{i=1}^{n} f(x_i)*f(y_i)}{\sqrt{\sum_{i=1}^{n} f^2(x_i)}*\sqrt{\sum_{i=1}^{n} f^2(y_i)}}$，当计算的相似度值越趋近于1，表明第

一文本和第二文本越相似，当计算的相似度值越趋近于-1，表明第一文本和第二文本越不相似。

[0064] 在一个实施例中，所述根据所述第一文本和第二文本获取对应的关键词库的步骤S2之前，包括：

[0065] S101：按照预设的规则将所述关键词划分为多个类别；

[0066] S102：将多个类别中的第一类关键词取出，并为其他关键词按照预设的权重规则划分权重；

[0067] S103：将多组相似文本依次输入至权重训练模型中进行训练，得到所述第一类关键词的权重参数。

[0068] 如上述步骤S101-S103所述，由于不同的第一文本或不同第二文本存在多个关键词，但是不同的关键词对于相似度的判定是不同的，应当为不同的关键词设置不同的权重，故而可以按照预设的规则将关键词划分为多个类别，由于第一类别关键词对相似度影响足够大，故而可以先为其他关键词按照预设的权重规则划分权重，然后将根据权重将多组相似文本向量化之后输入至权重训练模型中进行训练，具体可以为，将第一类别关键词的权重设置为参数，然后将多组相似文本向量化之后的向量按组依次输入至权重训练模型中，输出结果为该两个文本相似，训练完毕后得到第一类关键词的权重参数。

[0069] 在一个实施例中，所述基于所述关键词库提取所述第一文本的各第一关键词，并根据各所述第一关键词对应各第一权重值生成第一文本权重向量，以及，基于所述关键词库提取所述第二文本的各第二关键词，并根据各所述第二关键词对应各第二权重值生成第二文本权重向量的步骤S3，包括：

[0070] S301：通过文本分类器分别对所述第一文本和所述第二文本进行分词，分别对应得到第一词语和第二词语；

[0071] S302：根据所述关键词库将所述第一词语中对应的所述第一关键词，以及所述第二词语中对应的所述第二关键词提取出来；

[0072] S303：根据所述第一关键词对应各第一权重值生成第一文本权重向量，以及根据所述第二关键词对应各第二权重值生成第二文本权重向量。

[0073] 如上述步骤S301-S303所述，先通过文本分类器分别对第一文本和第二文本进行分词分别对应得到第一词语和第二词语，然后判断分词后的第一词语和第二词语中是否具有属于关键词库的关键词，将关键词库中具有的关键词，按照实现设定的关键词库中的规则进行设定每个关键词的权重，若不属于该关键词的权重，则可以将其设置为预设权重，由于其余不属于关键词的词语对第一文本和第二文本的相似度判定几乎没有什么作用，故而可以将其余不属于关键词的词语略去，不参与后续的计算过程，由此可以得到对应的第一文本权重向量和第二文本权重向量。

[0074] 在一个实施例中，所述根据所述第一文本和第二文本获取对应的关键词库的步骤S2，包括：

[0075] S201：将所述第一文本和所述第二文本依次输入至自然语言处理后的机器学习模型中，并计算得到所述第一文本与所述第二文本与各个类别分别对应的第一类别相似度和第二类别相似度；

[0076] S202：提取所述第一类别相似度中大于类别预设相似度的第一期望类别，以及提

取所述第二类别相似度中大于类别预设相似度的第二期望类别；

[0077] S203：提取所述第一期望类别中与所述第二期望类别中相同的目标类别，并获取所述目标类别对应的关键词库。

[0078] 如上述步骤S201-S203所述，由于不同的第一文本或第二文本具有不同的文本信息，因此其关键词库自然也对应不同，然后可以将第一文本和第二文本一次输入至自然语言处理后的机器学习模型中，对第一文本和第二文本的类别进行获取，具体获取的方法为通过在机器学习模型中预设多个类别，然后依次计算第一文本和第二文本的期望类别，由于第一文本和第二文本的相似情况未知，因此可以从第一期望类别和第二期望类别中选取相同的目标类别作为第一文本和第二文本的类别，然后根据目标类别获取对应的关键词库进行后续相似度的计算。

[0079] 在一个实施例中，所述根据所述第一文本和第二文本获取对应的关键词库的步骤S2之前，还包括：

[0080] S111：按照预设的规则将所述关键词划分为多个等级；

[0081] S112：设置最低等级的关键词的最低权重，以及根据公式$w_k = R_t + \sqrt{\sum_{t=k+1}^{c} n_t w_t^2}$设定其余等级的所述关键词的权重，其中，$w_c$表示所述最低权重，$w_t$表示第t等级的权重，$R_t$表示第t等级的预设参数，$n_t$表示第t等级所有关键词的总数量。

[0082] 如上述步骤S111-S112所述，先按照预设的规则将关键词划分为多个等级，即根据关键词的重要性划分不同的权重，该预设的规则可以是用户进行输入设定的，也可以是对第一文本和第二文本进行语义分析得到的，例如用户输入金融场景，则相应的等级可以包括保险业务关键词，应用场景关键词等，因此对保险业务关键词的权重系数大小的设定应当偏大一些，对应用场景关键词的权重系数大小的设定应当偏小一些，故而可以先设置最低等级的最低权重，然后按照公式依次设定其余等级的关键词的权重，应当理解的是，R_t的数值可以随着等级的变化而变化，也可以都是相同的参数。设置的目标权重应当满足$w_k > \sqrt{\sum_{t=k+1}^{c} n_t w_t^2}$，即需要大于语料库中优先级低于它的所有关键词的综合权重，即$R_t > 0$，从而实现对目标关键词设置目标权重，通过目标关键词来检测第一文本和第二文本的关键词，当然，目标权重也不宜设置过大，以免计算相似度时出现精度缺失。

[0083] 参照图2，本发明还提供了一种文本相似度的判断装置，包括：

[0084] 文本获取模块10，用于获取需要判断相似度的第一文本和第二文本；

[0085] 关键词库获取模块20，用于根据所述第一文本和第二文本获取对应的关键词库，其中，所述关键词库中存储有多个关键词，以及与各所述关键词一一对应的权重值；

[0086] 权重向量生成模块30，用于基于所述关键词库提取所述第一文本的各第一关键词，并根据各所述第一关键词对应各第一权重值生成第一文本权重向量，以及，基于所述关键词库提取所述第二文本的各第二关键词，并根据各所述第二关键词对应各第二权重值生成第二文本权重向量；

[0087] 相似度值计算模块40，用于根据相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值；

[0088] 相似判断模块50，用于根据所述相似度值判断所述第一文本与所述第二文本是否相似。

[0089] 在一个实施例中，相似度值计算模块40，包括：

[0090] 个数获取子模块，用于分别获取所述第一文本权重向量和所述第二文本权重向量中各关键词的个数；

[0091] 相似度值计算子模块，用于根据各关键词对应的所述权重以及个数，通过所述相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值；其中，所述相似度计算公式为$cos(I,R)=\frac{\sum_{i=1}^{n} f(x_i)*f(y_i)}{\sqrt{\sum_{i=1}^{n} f^2(x_i)}*\sqrt{\sum_{i=1}^{n} f^2(y_i)}}$，I表示所述第一文本权重向量，R表示第二文本权重向量，cos(I,R)表示所述相似度值，x_i表示所述第一文本权重向量的第i个关键词对应的个数，y_i表示所述第二文本权重向量的第i个关键词对应的个数，n表示所述关键词库中关键词的个数，$f(x_i)=\begin{cases}0, x=0\\ w_i, x>0\end{cases}$，$f(y_i)=\begin{cases}0, x=0\\ w_i, x>0\end{cases}$，$w_i$表示第个关键词对应的权重。

[0092] 在一个实施例中，文本相似度的判断装置，还包括：

[0093] 类别划分模块，用于按照预设的规则将所述关键词划分为多个类别；

[0094] 权重划分模块，用于将多个类别中的第一类关键词取出，并为其他关键词按照预设的权重规则划分权重；

[0095] 权重参数计算模块，用于将多组相似文本依次输入至权重训练模型中进行训练，得到所述第一类关键词的权重参数。

[0096] 在一个实施例中，权重向量生成模块30，包括：

[0097] 分词子模块，用于通过文本分类器分别对所述第一文本和所述第二文本进行分词，分别对应得到第一词语和第二词语；

[0098] 关键词提取子模块，用于根据所述关键词库将所述第一词语中对应的所述第一关键词，以及所述第二词语中对应的所述第二关键词提取出来；

[0099] 权重向量生成子模块，用于根据所述第一关键词对应各第一权重值生成第一文本权重向量，以及根据所述第二关键词对应各第二权重值生成第二文本权重向量。

[0100] 在一个实施例中，关键词库获取模块20，包括：

[0101] 类别相似度计算子模块，用于将所述第一文本和所述第二文本依次输入至自然语言处理后的机器学习模型中，并计算得到所述第一文本与所述第二文本与各个类别分别对应的第一类别相似度和第二类别相似度；

[0102] 提取子模块，用于提取所述第一类别相似度中大于类别预设相似度的第一期望类别，以及提取所述第二类别相似度中大于类别预设相似度的第二期望类别；

[0103] 目标类别提取子模块，用于提取所述第一期望类别中与所述第二期望类别中相同的目标类别，并获取所述目标类别对应的关键词库。

[0104] 在一个实施例中，文本相似度的判断装置，还包括：

[0105] 等级划分模块，用于按照预设的规则将所述关键词划分为多个等级；

[0106] 权重设定模块，用于设置最低等级的关键词的最低权重，以及根据公式

$w_k = R_t + \sqrt{\sum_{t=k+1}^{c} n_t w_t^2}$设定其余等级的所述关键词的权重，其中，$w_c$表示所述最低权重，$w_t$表示第t等级的权重，$R_t$表示第t等级的预设参数，$n_t$表示第t等级所有关键词的总数量。

[0107] 本发明的有益效果：通过为不同的关键词设置不同的权重，并且用权重来将第一文本和第二文本向量化，使其基于关键词的权重值计算相似度，从而可以提高第一文本和第二文本相似度判断与人为判断更接近。

[0108] 参照图3，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各种关键词库等。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可以实现上述任一实施例所述的文本相似度的判断方法。

[0109] 本领域技术人员可以理解，图3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。

[0110] 本申请实施例还提供一种计算机可读存储介质，其上存储有计算机程序，计算机程序被处理器执行时可以实现上述任一实施例所述的文本相似度的判断方法。

[0111] 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，所述的计算机程序可存储与一非易失性计算机可读取存储介质中，该计算机程序在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的和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。非易失性存储器可以包括只读存储器(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一多种形式可得，诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双速据率SDRAM(SSRSDRAM)、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。

[0112] 需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、装置、物品或者方法不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、装置、物品或者方法所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括该要素的过程、装置、物品或者方法中还存在另外的相同要素。

[0113] 以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的权利要求范围之内。

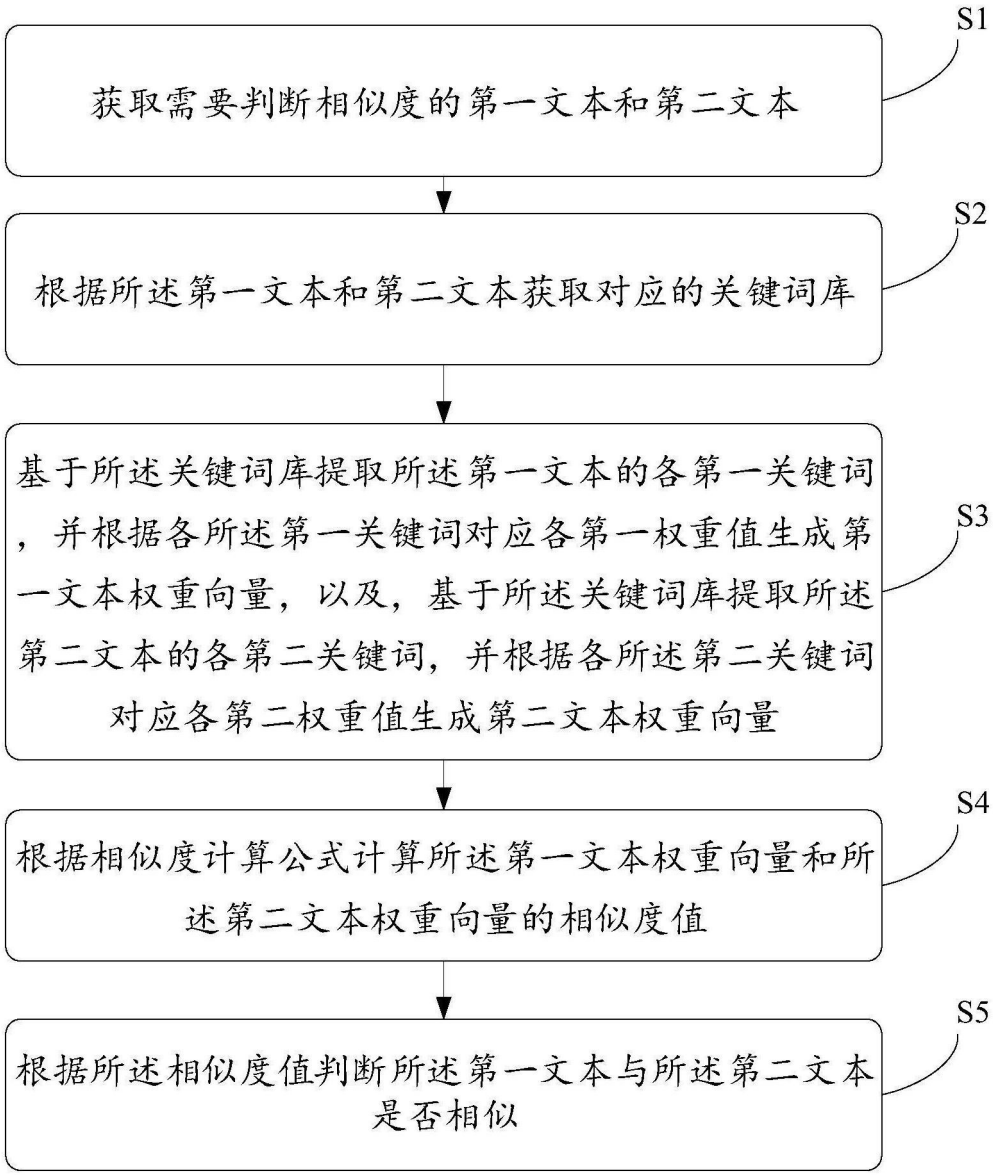

S1
获取需要判断相似度的第一文本和第二文本
S2
根据所述第一文本和第二文本获取对应的关键词库
S3
基于所述关键词库提取所述第一文本的各第一关键词，并根据各所述第一关键词对应各第一权重值生成第一文本权重向量，以及，基于所述关键词库提取所述第二文本的各第二关键词，并根据各所述第二关键词对应各第二权重值生成第二文本权重向量
S4
根据相似度计算公式计算所述第一文本权重向量和所述第二文本权重向量的相似度值
S5
根据所述相似度值判断所述第一文本与所述第二文本是否相似

图1

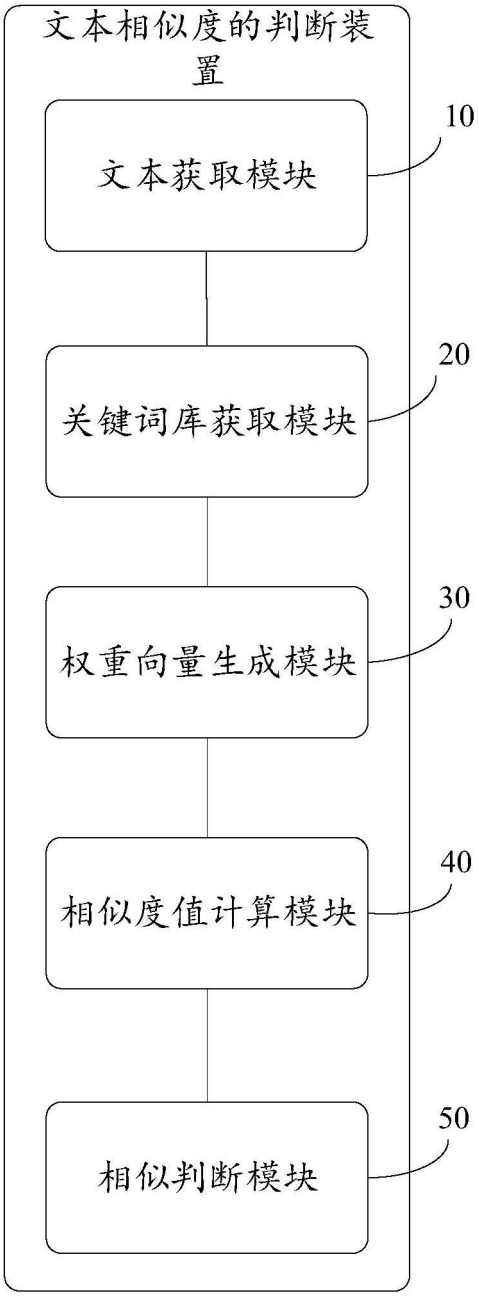

文本相似度的判断装置
文本获取模块
10
关键词库获取模块
20
权重向量生成模块
30
相似度值计算模块
40
相似判断模块
50

图2

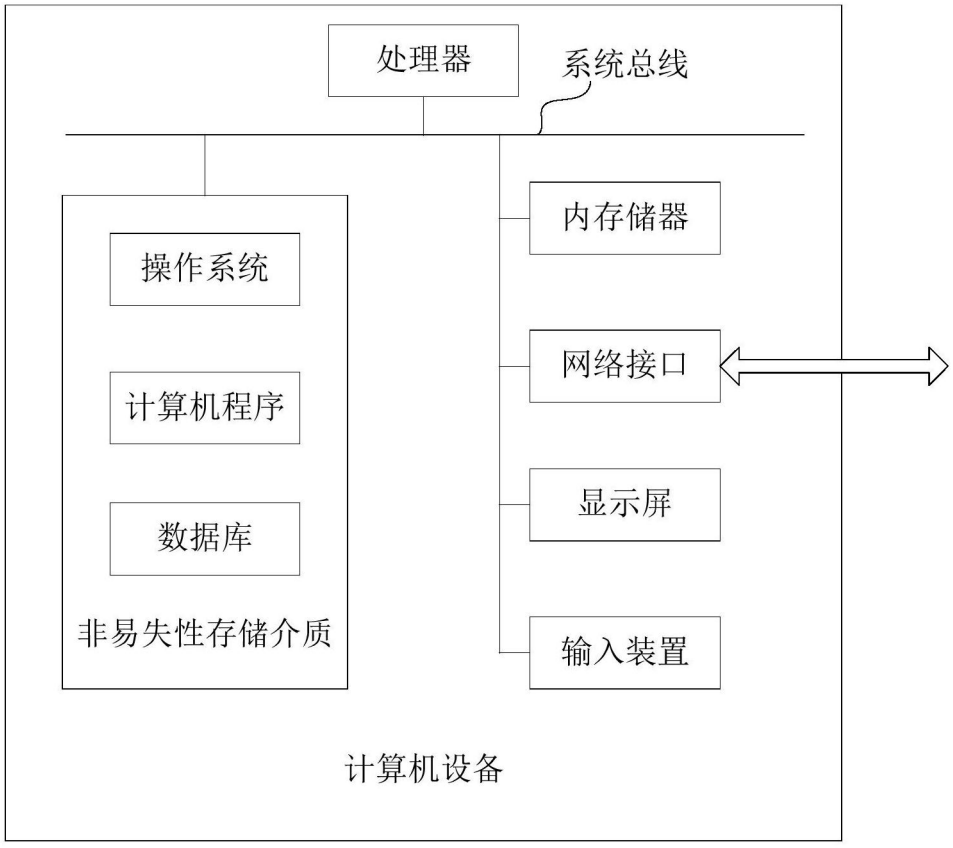


图3